



Nominalitzacions deverbales: Denotació y estructura argumental

Aina Peris Morant



Aquesta tesi doctoral està subjecta a la llicència *Reconeixement 3.0. Espanya de Creative Commons.*

Esta tesis doctoral está sujeta a la licencia *Reconocimiento 3.0. España de Creative Commons.*

This doctoral thesis is licensed under the *Creative Commons Attribution 3.0. Spain License.*

**NOMINALIZACIONES DEVERBALES:
DENOTACIÓN Y ESTRUCTURA
ARGUMENTAL**

AINA PERIS MORANT

Tesis presentada para optar
al grado de **Doctor en Lingüística** con mención europea
en el programa de doctorado *Ciencia Cognitiva y Lenguaje*,
Departamento de Lingüística,
Universidad de Barcelona,

bajo la supervisión de

Dra. Maria Taulé Delor

Universidad de Barcelona

Dr. Horacio Rodríguez Hontoria

Universidad Politécnica de Cataluña

Universidad de Barcelona

Febrero de 2012

I do not distinguish between the goals of theoretical and computational linguistics, but rather consider the use of computational tools and descriptions as an important part of the machinery for the analysis of linguistic theories.

James Pustejovsky
Generative Lexicon (1995:40).

A mis padres, por estar siempre a mi lado.

RESUMEN

Las nominalizaciones deverbales del español son construcciones lingüísticas que se caracterizan por presentar propiedades propias de los sustantivos pero al mismo tiempo poder heredar la estructura argumental de los verbos de los que derivan. Esta dualidad les confiere un notable interés lingüístico porque, por una parte, pueden denotar tanto un estado o el resultado de la acción denotada por el verbo base correspondiente, como pueden también 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. Además del intrínseco valor lingüístico que tiene el estudio de estas construcciones, también desde un punto de vista del Procesamiento del Lenguaje Natural 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 coreferencia 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 de búsqueda de respuestas, así como los sistemas de etiquetado semántico, pueden aprovechar herramientas y recursos que representen la estructura argumental de las nominalizaciones deverbales.

Esta tesis pretende conjugar el estudio de las nominalizaciones deverbales tanto desde un punto de vista lingüístico como desde la perspectiva del Procesamiento del Lenguaje Natural. La tesis está dividida en cuatro partes que responden a esa voluntad.

La primera parte de este trabajo nos pone en antecedentes acerca de las nominalizaciones deverbales. Se define el objeto de estudio, se presenta la metodología utilizada y se ofrece una revisión bibliográfica amplia que incluye tanto trabajos fundamentalmente teóricos como trabajos esencialmente computacionales sobre las nominalizaciones deverbales.

La segunda parte se centra en la estructura argumental de las nominalizaciones deverbales. En primer lugar, se presenta el estudio lingüístico basado en corpus sobre la realización sintáctico-semántica de los argumentos. A partir de este estudio, se extraen una serie de hipótesis lingüísticas sobre qué constituyentes son argumentos de las nominalizaciones y cuáles no, y qué tipo de argumento verbal se asocia a constituyentes específicos en el dominio nominal. En segundo lugar, estas hipótesis lingüísticas están en la base del paquete de reglas heurísticas (RHN) creado para anotar automáticamente la estructura argumental de las nominalizaciones deverbales en el corpus AnCora-Es. La evaluación de estas reglas heurísticas aporta nuevas observaciones sobre la realización de la estructura argumental de las nominalizaciones deverbales y confirma parte de las hipótesis iniciales.

La tercera parte trata sobre la denotación de las nominalizaciones deverbales. Primero, se presenta el estudio empírico basado en corpus realizado sobre la distinción entre evento y resultado. De este estudio empírico se obtienen una serie de criterios lingüísticos para establecer dicha distinción, y además, se establece una nueva clase denotativa subespecificada para aquellos casos en los que el contexto oracional es insuficiente. Los criterios lingüísticos resultan de la determinación de qué criterios propuestos en la bibliografía son relevantes para el español, del análisis lingüístico realizado, y de la observación de las reglas simbólicas generadas en los experimentos computacionales para evaluar los criterios anteriores. Estos experimentos están en la base del clasificador ADN, un sistema automático cuyo objetivo es clasificar las nominalizaciones deverbales según su denotación. Este clasificador se desarrolló como herramienta necesaria para la anotación de la denotación de las nominalizaciones deverbales del corpus AnCora-Es y, finalmente, se ha convertido en el primer clasificador de denotaciones del español capaz de trabajar en diferentes escenarios.

En la cuarta parte se describen los dos recursos generados en esta tesis: el enriquecimiento del corpus AnCora-Es con la anotación de la denotación y la estructura argumental de las nominalizaciones deverbales, y la inducción del léxico AnCora-Nom a partir de esta anotación. En relación a AnCora-Es, se detallan los procesos de validación manual de la estructura argumental y la denotación, concretamente, los criterios específicos de validación y las pruebas de acuerdo entre anotadores. Respecto a AnCora-Nom, se especifica la generación automática del léxico a partir del corpus validado, evidenciando la posibilidad de obtener dos recursos con un único proceso de validación manual, el del corpus.

Finalmente, en las conclusiones se recogen las aportaciones de esta tesis a la comunidad científica. Estas aportaciones consisten básicamente en herramientas y recursos computacionales para el tratamiento y representación de las nominalizaciones deverbales del español, y en el análisis lingüístico que caracterizan las nominalizaciones deverbales tanto con respecto a la denotación como a la estructura argumental, conjugando las dos perspectivas de estudio de este trabajo.

ABSTRACT

Spanish deverbal nominalizations are linguistic constructions characterized by presenting properties of common nouns but also by inheriting the argument structure of the verbs from which they derive. This duality aroused considerable interest in deverbal nominalizations in Linguistics. On the one hand, they can denote both the state or the result of the action expressed by the corresponding base verb as well as the same action or event expressed by the base verb, in the latter being paraphrases of sentence clauses. On the other hand, nominalizations are nouns with argument taking capacity, that is, they select arguments. In this sense, it is relevant to observe the patterns of the syntactic-semantic realization of the nominalizations arguments, since they represent an alternative way of expressing the same semantic content of a sentence.

Besides the intrinsic linguistic value of studying these constructions in Spanish, having tools and resources dealing with deverbal nominalizations is essential in Natural Language Processing (NLP), both in terms of denotation and argument structure. Tasks such as coreference resolution or paraphrase detection may benefit from a tool or resource that addresses the denotation type of nominalizations. Applications such as information extraction or question answering, and semantic role labelling systems may also benefit from tools and resources that represent the argument structure of deverbal nominalizations.

This thesis aims to study deverbal nominalizations both from Linguistics and NLP approaches. The thesis is divided into four parts, which reflect these two perspectives.

The first part of this work gives background information on deverbal nominalizations. It defines the object of study, presents the methodology used and provides an extensive review of the literature, including both theoretical and computational works on deverbal nominalizations.

The second part focuses on the argument structure of deverbal nominalizations. We present our corpus-based linguistic study of the syntactic-semantic realization of arguments. From this study, we extracted a series of hypotheses about which constituents are arguments of nominalizations and which are not, and what

kind of verbal argument is associated with specific constituents in the nominal domain. These assumptions underlie the RHN package of heuristics rules created to automatically annotate the argument structure of deverbal nominalizations in the Ancora-Es corpus. The evaluation of these heuristics provides new observations on the realization of the argument structure of deverbal nominalizations and confirms part of our initial hypotheses.

The third part deals with the denotation of deverbal nominalizations. First, we present our empirical corpus-based study of the distinction between event and result nominalizations. From this empirical study a series of linguistic criteria for establishing that distinction was obtained. We also established a new denotative class, underspecified, for those cases in which the sentence context is not enough for disambiguation. The linguistic criteria result from determining which criteria proposed in the literature are relevant for Spanish, from the linguistic analysis performed, and from the observance of the symbolic rules generated in the computational experiments to evaluate the above criteria. These experiments are in the base of the ADN-Classifier, an automatic system for the classification of deverbal nominalizations according to their denotation. This classifier was developed as a necessary tool for annotating the denotation of deverbal nominalizations in the Ancora-Es corpus and it has become the first tool for the automatic classification of deverbal nominalizations into denotation types that can work in different scenarios.

The fourth part describes the two resources generated in this thesis: the enrichment of the Ancora-Es corpus by annotating the denotation and argument structure of deverbal nominalizations, and the extraction from this annotation of the Ancora-Nom lexicon. Regarding Ancora-Es, we detail the manual validation processes of the argument structure and denotation, namely, specific validation criteria and inter-annotator agreement tests. Regarding Ancora-Nom, we specify the automatic generation of the lexicon from the validated corpus, demonstrating the possibility of obtaining two resources with a single manual validation process of the corpus.

Finally, the contributions of this thesis to the scientific community are presented in the conclusions. These contributions consist of, on the one hand, computational tools and resources for the treatment and representation of Spanish deverbal nominalizations. And, on the other hand, the linguistic analysis carried out to characterize deverbal nominalizations with respect to both their denotation and their argument structure, combining the two approaches of this work.

AGRADECIMIENTOS

Para llevar a cabo esta tesis he contado con el apoyo de muchas personas, pero esta no habría sido posible sin mis directores de tesis, Mariona Taulé Delor y Horacio Rodríguez Hontoria, quienes, además de transmitirme su vocación investigadora, me han orientado, ayudado y animado constante y directamente en todos los aspectos de la tesis durante estos cuatro años. Agradecerles la confianza que siempre me han demostrado, así como la dedicación y la atención que en todo momento me han ofrecido. A Mariona, además, le doy las gracias por la paciencia, la vitalidad y el ánimo que siempre me transmite.

Mi más sincera gratitud también a Maria Antònia Martí, por su más que generoso apoyo para la realización de esta tesis, tanto desde un punto logístico como personal. Valoro especialmente el ímpetu y la pasión investigadora que demuestra cada día y sobre todo, que siempre encuentre un hueco en su apretada agenda para escucharte y aconsejarte respecto a cualquier asunto que sea objeto de preocupación.

Esta tesis también se ha beneficiado de las personas que he encontrado en mis dos estancias en el extranjero. Al *Institut für Maschinelle Sprachverarbeitung* de la Universidad de Stuttgart (Alemania) acudí bajo la supervisión de Ulrich Heid, quien dirigía un magnífico grupo formado por Gertrud Faasz, Kati Schweitzer, Ekaterina Lapshinova-Koltunski, Kurt Eberle y Kerstin Eckart. A todos ellos les doy las gracias por acogerme tan bien y por sus ganas interminables de debatir sobre la semántica de las nominalizaciones. Mi estancia en el *Computer Science Department* de la Universidad de Nueva York fue dirigida por Adam Meyers, a quien le agradezco su amabilidad y que compartiera conmigo toda la experiencia adquirida sobre las nominalizaciones deverbales en el proyecto NomBank. Al resto de miembros de *The Proteus Project* les doy las gracias por las interesantísimas reuniones de los martes, especialmente a Cristina Mota, Xu Wei y Bonan Min por sus enriquecedoras sugerencias sobre mi trabajo.

De vuelta a Barcelona, mi agradecimiento se dirige a todas las personas que conforman el Departamento de Lingüística General de la Universidad de Barcelona y, especialmente, el grupo de investigación CLiC, que contribuyen a un

excelente clima de trabajo y que siempre están dispuestas a echar una mano. Una mención especial para los que han participado como anotadores manuales de las nominalizaciones deverbales, porque sin su trabajo esta tesis no sería posible. Gracias a todos: Esther Arias, Oriol Borrega, Santiago González, Difda Monterde, Lourdes Puiggròs y Rita Zaragoza.

No puedo olvidar a Manu Bertran, el informático del grupo, que nos hace el trabajo mucho más sencillo, ni por supuesto a David Bridgewater, por ser mucho más que un profesor de inglés. Tampoco se me pueden pasar por alto todos los becarios de CLiC, Glòria de Valdívía, Raquel Garrido, John Roberto, Marta Vila y Marta Recasens, con los que he compartido intereses y preocupaciones a partes iguales. Un especial y afectuoso agradecimiento a las Martas, por dejarme ser la Z en el mundo de las ecuaciones.

Finalmente, también quiero dar las gracias a todas las personas que desde fuera del mundo académico han contribuido a que realizara esta tesis: los amigos y la familia. A los amigos, porque sin los momentos compartidos con ellos las fuerzas no serían las mismas. En concreto, agradezco al sector Calabria (Marta, Belén, Nadia, Elena y Jana) sus altas dosis de cariño y buen humor; a Cice, su particular mirada sobre los problemas; a Sílvia, nuestras charlas revitalizadoras; y a Marina, aquella magdalena de chocolate en horas bajas cuyo recuerdo me acompaña siempre. A mi familia le agradezco su comprensión y afecto incondicionales. A mi hermano le doy las gracias por los abrazos voladores que me llenan de energía, y a mis padres, por ser siempre mi mejor y más seguro sostén, una fuente de tranquilidad y confianza esencial para mí. A Juan, mi pareja, le doy las gracias por haber llegado a mi vida en la época del doctorado, pero, sobre todo, por quedarse.

Esta tesis ha sido financiada por una beca (AP2007-01028) del Ministerio de Educación del Gobierno de España.

ÍNDICE GENERAL

Resumen	VII
Abstract	IX
Agradecimientos	XI
Índice general	XIII
Índice de figuras	XVII
Índice de tablas	XIX
Índice de acrónimos	XXI
I Antecedentes	1
1. Introducción	3
1.1. La necesidad de estudiar las nominalizaciones para el PLN	7
1.1.1. ¿Qué nominalizaciones deverbales estudiamos?	7
1.1.2. La importancia para el PLN de las nominalizaciones	9
1.2. Objetivos del trabajo	11
1.3. Procedimiento	12
1.4. Contribuciones	16
1.5. Estructura de la tesis	17

2. Nominalizaciones deverbales: estado de la cuestión	21
2.1. Aproximaciones lingüísticas	21
2.1.1. Nominalizaciones deverbales y denotación	22
2.1.2. Nominalizaciones deverbales y estructura argumental	36
2.2. Aproximaciones Computacionales	41
2.2.1. Recursos	41
2.2.2. Sistemas	48
II Estructura Argumental	59
3. Estructura argumental de las nominalizaciones deverbales: estudio empírico	61
3.1. Extracción de la muestra de datos	62
3.2. Esquema de anotación	63
3.3. Estructura argumental: análisis lingüístico	66
3.4. Conclusiones	71
4. Anotación automática de los argumentos internos	73
4.1. Reglas Heurísticas y Recursos Lingüísticos	73
4.1.1. Recursos Lingüísticos	75
4.1.2. Reglas Heurísticas	77
4.2. Evaluación de la anotación automática de la estructura argumental	92
4.3. Discusión	98
4.3.1. Comparación de resultados	100
4.4. Conclusiones	102
III Denotación	103
5. La denotación en las nominalizaciones deverbales: estudio empírico	105
5.1. Denotación: análisis lingüístico	105
5.1.1. Análisis de los criterios de la bibliografía	107
5.1.2. Nuevos indicadores de la denotación	114
5.2. Denotación: análisis computacional	116
5.2.1. Experimentos para la evaluación de AnCora-Nom-v1	118
5.2.2. Criterios a partir de la observación de las reglas del modelo de clasificación	127
5.3. Conclusiones	129

6. Clasificador ADN	131
6.1. Clasificador ADN	132
6.2. Rasgos utilizados y recursos lingüísticos	137
6.2.1. Rasgos obtenidos de AnCora-Nom	137
6.2.2. Rasgos obtenidos del corpus AnCora-Es	138
6.2.3. Rasgos obtenidos del léxico AnCora-Verb	138
6.3. Conclusiones	139
7. Clasificador ADN: experimentos	141
7.1. Marco de desarrollo	141
7.2. Experimentos	142
7.3. Evaluación	144
7.3.1. Clasificador orientado a la precisión	147
7.3.2. Evaluación de los escenarios	148
7.3.3. Análisis de errores	150
7.4. Discusión	155
7.5. Conclusiones	156
IV Recursos	157
8. AnCora-Es: validación manual	159
8.1. Validación manual de la estructura argumental	159
8.1.1. Descripción de la tarea de validación manual	160
8.1.2. Criterios de anotación	162
8.1.3. Pruebas de acuerdo entre anotadores	165
8.2. Validación manual de la denotación	167
8.2.1. Descripción de la tarea de validación manual	168
8.2.2. Criterios lingüísticos para la clasificación de las nominalizaciones deverbales según su denotación	170
8.2.3. Pruebas de acuerdo entre anotadores	176
8.3. Adaptación de AnCora-Pipe para la anotación de los SNs	177
8.4. Conclusiones: AnCora-Es-v3	184
9. AnCora-Nom: un léxico de nominalizaciones deverbales	187
9.1. Proceso de creación del léxico AnCora-Nom	187
9.1.1. Proceso de extracción	189
9.2. AnCora-Nom	196
9.2.1. Atributos a nivel de entrada léxica	197
9.2.2. Atributos a nivel de sentido	198
9.2.3. Atributos a nivel de marco	202

9.3. Análisis cuantitativo de los datos	206
9.4. Conclusiones	210
 ***	 213
10. Conclusions and Further Work	215
10.1. Contributions	215
10.1.1. Linguistic Findings	216
10.1.2. Tools	219
10.1.3. Lexical resources	220
10.2. Further Work	221
10.2.1. Immediate work	221
10.2.2. Future work	227
 Bibliografía	 229
 Apéndices	 243
A. Lista de adjetivos relacionales	245
B. Lista de publicaciones relacionadas con la tesis	247

ÍNDICE DE FIGURAS

1.1.	Esquema del procedimiento utilizado en el desarrollo de la tesis . . .	13
4.1.	Proceso de anotación de la estructura argumental	75
4.2.	Frecuencia de las combinaciones de constituyentes en los SNs . . .	85
5.1.	Esquema de los experimentos computacionales para la verificación de los criterios	119
6.1.	Funcionamiento del Clasificador ADN	133
6.2.	Árbol sintáctico parcial que contiene la nominalización ‘aumento’	139
7.1.	Curva de aprendizaje para el modelo LEAFF	148
7.2.	Cobertura y precisión para el modelo LEAFF.	149
8.1.	Validación manual de la estructura argumental	161
8.2.	Entrada léxica del verbo ‘volar’ en AnCora-Verb	164
8.3.	Validación manual de la denotación	169
8.4.	Aplicación de los criterios para la distinción Evento vs. Resultado	175
8.5.	AnCora-Pipe para la anotación de los SNs.1	179
8.6.	AnCora-Pipe para la anotación de los SNs.2	181
8.7.	AnCora-Pipe para la anotación de los SNs.3	182
8.8.	AnCora-Pipe para la anotación de los SNs.4	183
8.9.	Ejemplo de anotación de ‘ampliación’ en AnCora-Es	184
9.1.	Proceso de elaboración incremental del léxico AnCora-Nom	190
9.2.	Estructura de entrada léxica de AnCora-Nom	191
9.3.	Entrada léxica de ‘aceptación’	198
9.4.	Entrada léxica del sentido lexicalizado ‘golpe de estado’	201

10.1. File in the corpus with the nominalization ‘decisión’	224
10.2. Syntactic structure of sentence (4)	225

ÍNDICE DE TABLAS

2.1. Tabla resumen de las clasificaciones según la denotación de las nominalizaciones deverbales	31
2.2. Criterios Lingüísticos para la distinción Evento vs. Resultado . . .	33
2.3. Recursos lingüísticos que representan las nominalizaciones deverbales	47
2.4. Sistemas automáticos para el tratamiento computacional de las nominalizaciones deverbales	58
3.1. Conjunto de etiquetas argumentales utilizadas en la anotación de las nominalizaciones deverbales	65
4.1. Clases semánticas verbales	76
4.2. Notación simplificada de las Reglas Generales	78
4.3. Notación simplificada de las reglas específicas de un constituyente	82
4.4. Correspondencia entre la clase semántica verbal, argumentos y papeles temáticos	84
4.5. Notación Simplificada de las reglas específicas de dos SPs	86
4.6. Notación Simplificada de las reglas específicas de Poss+SP/SA . .	87
4.7. Notación Simplificada de las reglas específicas de SP + SA	88
4.8. Notación Simplificada de las reglas específicas de dos SAs	90
4.9. Notación Simplificada de las reglas específicas de GRel+SP/SA .	91
4.10. Resultados de la anotación automática por constituyentes	92
4.11. Resultados de la anotación automática por constituyentes y etiquetas	93
4.12. Eficacia de las reglas generales para los SPs	97
5.1. Resultados de los criterios por denotaciones	111

5.2.	Rasgos utilizados en los experimentos a nivel de sentido para la validación empírica de AnCora-Nom-v1	120
5.3.	Resultados de los experimentos a nivel de sentido para la validación empírica de AnCora-Nom-v1	121
5.4.	Análisis de errores de los experimentos a nivel de sentido para la validación empírica de AnCora-Nom-v1	123
5.5.	Matriz de confusión de los experimentos a nivel de sentido para la validación empírica de AnCora-Nom-v1	123
5.6.	Rasgos contextuales empleados en los experimentos a nivel de corpus.	124
5.7.	Resultados de los experimentos a nivel de sentido añadiendo rasgos de AnCora-Es a los rasgos de AnCora-Nom-v1	126
5.8.	Tipo denotativo según la realización argumental de nominalizaciones derivadas de verbos de la clase semántica de los logros . .	128
5.9.	Tipo denotativo según la realización argumental de nominalizaciones derivadas de verbos de la clase semántica de las realizaciones	130
6.1.	Contenido descriptivo de AnCora-Es.	134
6.2.	Escenarios	135
7.1.	Experimentos y Evaluación de los modelos.	146
7.2.	Experimentos y evaluación de los escenarios	149
7.3.	Matriz de confusión del modelo LEAFF	150
8.1.	Resultados de la prueba de acuerdo entre anotadores: estructura argumental	167
8.2.	Resultados de la prueba de acuerdo entre anotadores: denotación .	176
9.1.	Distribución de los sentidos nominales: denotación y lexicalización	207
9.2.	Distribución de los sentidos nominales: denotación y número de argumentos	208
9.3.	Distribución de los sentidos nominales: denotación y tipo de determinante	209
9.4.	Distribución de los distintos tipos de argumentos según el tipo de constituyente	210

ÍNDICE DE ACRÓNIMOS

- ADN: Anotación la Denotación de las Nominalizaciones deverbales, 11
- CDT: *Copenhaguen Dependency Treebank*, 46
- EP: Estructura Profunda, 36
- ES: Estructura Superficial, 36
- FUDRS: *Flat Underspecified Discourse Representation Structures* (Estructuras de representación del discursos llanas y subespecificada), 49
- GB: *Government and Binding Theory* (Teoría de la Rección y el Ligamento), 36
- GG: *Generative Grammar* (Gramática Generativa), 23
- GL: *Generative Lexicon* (Lexicón Generativo), 26
- GRel: pronombre Genitivo Relativo, 66
- HPSG: *Head-driven Phrase Structure Grammar* (Gramática de Estructura Sintagmática regida por el Núcleo), 36
- LFG: *Lexical Functional Grammar* (Gramática Léxico-Funcional), 36
- ML: *Machine Learning* (Aprendizaje Automático), 14
- MTT: *Meaning-Text Theory* (Teoría Sentido-Texto), 26
- NC: Nombre Común, 62
- OSub: Oración Subordinada, 66
- PLN: Procesamiento del Lenguaje Natural, 5
- PoS: *Part of Speech* (categoría morfológica), 53
- Poss: determinate POSeSivo, 66
- RHN: Reglas Heurísticas para las Nominalizaciones deverbales, 11
- RRG: *Role Reference Grammar* (Gramática del Rol y la Referencia), 30
- SA: Sintagma Adjettival, 66

SAdv: Sintagma Adverbial, 66

SN: Sintagma Nominal, 4

SP: Sintagma Preposicional, 4

SRL: *Semantic Role Labeling* (Etiquetado de Roles Semánticos), 6

SVM: *Support Vector Machine* (máquinas de vectores de soporte), 50

WSD: *Word Sense Disambiguation* (Desambiguación de Sentidos), 5

Parte I

Antecedentes

CAPÍTULO 1

INTRODUCCIÓN

Las lenguas disponen de múltiples mecanismos para expresar conceptos similares, aunque también es cierto que cada posibilidad supone un matiz distinto en el significado expresado. Esta versatilidad refleja la riqueza del lenguaje, que tanto nos fascina y nos interesa estudiar. Concretamente desde la Lingüística Computacional, área en la que se enmarca este trabajo, se trata de modelar el lenguaje natural para que pueda ser procesado computacionalmente. En este sentido, dicho modelado no puede hacerse sino parcialmente, es decir, atendiendo a parcelas, entiéndase, construcciones concretas del lenguaje. Esta tesis se centra en las nominalizaciones deverbales del español, una construcción lingüística que encierra un importante contenido semántico pero que, sin embargo, no ha sido estudiada en el ámbito computacional hasta hace poco porque la mayor parte de la atención se ha centrado en el verbo. A pesar de que la misma cronología se ha dado también en inglés, es decir, se ha prestado atención con anterioridad a los verbos que a las nominalizaciones, también es cierto que en esta lengua a partir de los años 90 empiezan a aparecer ya trabajos relevantes que toman como foco de estudio estas construcciones (Hindle, 1990; Macleod et al., 1998).

- (1) [La patronal]x **propone** [**ampliar** [de ocho a doce meses]z [el periodo de referencia para poder solicitar el subsidio de desempleo]w]y¹.
- (2) La **propuesta** [de la patronal]x [de instaurar la **ampliación** [de ocho a doce meses]z [del periodo de referencia para poder solicitar el subsidio de desempleo]w]y se ha aceptado.

¹Todos los ejemplos, excepto los contrariamente indicados, se han obtenido del corpus AnCora-Es (Taulé et al., 2008).

Fijémonos en los ejemplos (1) y (2). ¿Qué diferencia existe entre ambas oraciones? ¿Qué tipo de significado transmiten? ¿Se trata de informaciones distintas? Si observamos con atención ambas oraciones nos damos cuenta que las dos expresan el mismo contenido semántico, es decir, se trata de significados equivalentes que hacen referencia al mismo proceso o representación mental (Recasens and Vila, 2010). Una representación formal del tipo “X-agente *evento* Y-paciente”, por ejemplo, sirve tanto para representar el significado de la oración (1) como de la (2), lo que varía es la construcción sintáctica mediante la cual se expresa el evento. En la primera oración el evento principal se expresa mediante un predicado verbal (‘proponer’), mientras que en la segunda oración es un predicado nominal (‘propuesta’) el que expresa el mismo evento. Evidentemente, esto tiene consecuencias a nivel sintáctico en la manera en que se expresan los argumentos. En la oración (1) el argumento agente (X) se realiza mediante un sintagma nominal (SN, en adelante) con la función sintáctica de sujeto (‘la patronal’) y el argumento paciente (Y) se realiza mediante una oración subordinada de infinitivo (‘ampliar de ocho a doce meses...’), que funciona como complemento directo en esta construcción transitiva. En el ejemplo (2) los argumentos agente (X) y paciente (Y) se realizan mediante sintagmas preposicionales (SPs, en adelante) (‘de la patronal’ y ‘de instaurar la ampliación...’, respectivamente) puesto que son complementos del nombre ‘propuesta’.

Si observamos los ejemplos (1) y (2), nos damos cuenta de que además del evento principal, también el evento que se describe en la oración subordinada completiva del ejemplo (1) (‘ampliar de ocho a doce meses...’) aparece en forma de predicado nominal en el ejemplo (2) (‘ampliación’), y ambos predicados tienen los mismos argumentos, un argumento paciente (marcado por el índice W en cada ejemplo) y un argumento extensión (marcado por el índice Z) aunque, como antes, la realización sintáctica de dichos argumentos difiere: el argumento paciente se realiza mediante un SN (‘el periodo de referencia...’) en el ejemplo (1) y como SP (‘del periodo de referencia...’) en el ejemplo (2).

Por lo tanto, se trata de dos maneras alternativas –predicado verbal vs. predicado nominal– de expresar un mismo evento. De modo que si queremos analizar el contenido semántico de un texto, si estamos interesados en el estudio y representación del significado, en analizar qué tipo de relaciones se establecen entre los predicados y sus argumentos, tenemos que contemplar también los predicados nominales. Es precisamente este hecho el que nos planteó la necesidad de estudiar las nominalizaciones deverbales y su representación semántica en español ya que transmiten importante contenido semántico. Además, dado que son relativamente frecuentes en el lenguaje escrito esta necesidad resultó si cabe, más patente. En Hull and Gomez (2000, p.141-142) nos dicen que de cada 25 párrafos seleccionados aleatoriamente de la *World Book Encyclopedia*, en 23 aparecen al menos 2 nominalizaciones en cada uno de ellos.

Importancia de las
nominalizaciones
deverbales

En el corpus AnCora-Es (Taulé et al., 2008) aparecen 23.431 nominalizaciones deverbales que suponen aproximadamente el 30 % de los predicados que codifican la información semántica del corpus; el 70 % restante son predicados verbales (56.590 ocurrencias). Todo esto demuestra que las nominalizaciones deverbales son construcciones que se utilizan asiduamente para expresar importantes contenidos semánticos, por lo que no tenerlas en cuenta constituye un error.

Siguiendo con los ejemplos (1) y (2), imaginemos un Sistema de Búsqueda de Respuestas al que se le hacen las siguientes preguntas: ‘¿Qué ha propuesto la patronal?’, ‘¿En cuánto se amplía el periodo de referencia?’, ‘¿Quién quiere ampliar el periodo de referencia?’ Si solo disponemos de la información representada en (2), y solo tuviéramos analizados y representados los predicados verbales, no podríamos obtener ninguna respuesta a las preguntas anteriores. En este caso, el único predicado verbal es ‘aceptar’ y, por lo tanto, su representación no ofrece ninguna respuesta para las cuestiones planteadas. Esto demuestra que no contemplar los predicados nominales, como en este caso, realmente supone una pérdida de información. Este trabajo pretende ampliar la capacidad de búsqueda de estos sistemas, estudiando las nominalizaciones deverbales del español en el marco de la Lingüística Computacional.

Esta disciplina ha trabajado recientemente de forma intensa en el tratamiento semántico de textos no restringidos. Son una clara muestra la Semántica Recursiva Mínima de Lingo/LKB *Minimal Recursive Semantics in Lingo/LKB*, (Copestake, 2007), la Semántica de Marcos, *Frame Semantics* utilizada en Shalmaneser (Erk and Padó, 2006), las Estructuras de Representación del Discurso *Discourse Representation Structures* presentes en la herramienta Boxer (Bos, 2008) o el aprendizaje automático de las Gramáticas Semánticas, *Semantic Grammars* (Mooney, 2007). Sin embargo, aún se está lejos de representar completamente el significado de los textos si no se restringen a dominios concretos. Además, muchas aplicaciones del Procesamiento del Lenguaje Natural (en adelante, PLN) como son la Extracción de Información, los Sistemas de Búsqueda de Respuestas, la Lectura Automática (*Machine Reading*), la Traducción Automática y tareas de nivel intermedio como la Implicación Textual (*Textual Entailment*), la Detección de Paráfrasis o la Desambiguación de Sentidos (*Word Sense Disambiguation*, WSD) han alcanzado sus cotas reales superiores con las aproximaciones que actualmente se siguen y no pueden ser mejoradas sino es mediante el uso de una representación semántica adecuada del texto en cuestión.

Dadas las limitaciones y dificultades en obtener de forma automática una representación semántica profunda de los textos, los esfuerzos se han dirigido a representaciones semánticas parciales que usan formalismos semánticos menos expresivos (a menudo variantes de la Lógica de Descripciones (*Description Logic*) o se ha descartado la posibilidad de representar el texto en su conjunto para centrarse en tareas más sencillas. Este es el caso de los sistemas de Etiquetado de

Roles Semánticos (*Semantic Role Labeling, SRL*), que indican qué tipo de relaciones semánticas mantiene un predicado con sus participantes correspondientes siendo estas relaciones obtenidas a partir de una lista predefinida de posibles papeles temáticos para un predicado o clase de predicado dados. Véase Márquez et al. (2008) y Palmer et al. (2010) para revisiones recientes de este tipo de sistemas. Estrechamente relacionada con el SRL se encuentra la tarea de aprendizaje de Restricciones de Selección (*Selectional Restrictions*) para un predicado. Esta tarea consiste en aprender la clase semántica a la que pertenece cada argumento de un predicado (Mechura, 2008). También en este caso se utiliza un conjunto predefinido de etiquetas semánticas para llevar a cabo la tarea de clasificación. WordNet (Fellbaum, 1998) es uno de los recursos más utilizados para este fin. Con todo, la mayor parte de estos esfuerzos se han centrado principalmente en el verbo, considerado, en general, el núcleo de la oración, el elemento vertebrador del significado, relegando a un segundo plano otros tipos de predicados como, por ejemplo, las nominalizaciones deverbales que, como muestra el ejemplo (2), también son construcciones equivalentes para expresar un evento. Conscientes de dicha limitación, recientemente ha surgido un interés en ir más allá del verbo en el tratamiento semántico de textos. En esta línea encontramos los trabajos desarrollados por Meyers (2007), Ruppenhofer et al. (2006), Lapata (2002), Girju et al. (2009), Padó et al. (2008) y en *The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies* Surdeanu et al. (2008), entre otros, que se han centrado en el tratamiento y representación semántica de las nominalizaciones deverbales, básicamente para el inglés. En este sentido, son pocos los trabajos que tratan las nominalizaciones deverbales en otras lenguas. En el proyecto FrameNet (Ruppenhofer et al., 2006), además del inglés, se representan las nominalizaciones deverbales del alemán (Burchardt et al., 2009), el japonés (Ohara, 2009) o el español (Subirats, 2009), aunque el número de nominalizaciones representadas es inferior en estas lenguas respecto al contenido del FrameNet inglés. En francés encontramos el trabajo que están desarrollando en el grupo Nomage (Balvet et al., 2010), para el ruso existe “The Essex Database of Russian Verbs and Their Nominalizations” (Spencer and Zaretskaya, 1999) y para el chino se han desarrollado sistemas de SRL (Xue, 2006).

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 ha parecido 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 creemos necesario emprender este estudio en español con el objetivo de dotar a esta lengua de herramientas y recursos para su tratamiento computacional.

Este capítulo se organiza en cinco secciones. En la primera se detallan las razones que desde el PLN emergen para que las nominalizaciones deverbales sean estudiadas (Sección 1.1). En la segunda se definen los objetivos de la tesis (Sección 1.2), en la tercera se describe el procedimiento seguido para desarrollar esta investigación (Sección 1.3), en la cuarta se adelantan las contribuciones de este trabajo (Sección 1.4) y finalmente, en la última sección se presenta la estructura organizativa de esta tesis (Sección 1.5).

1.1. La necesidad de estudiar las nominalizaciones deverbales del español para el PLN

En esta sección se explica porqué el estudio de las nominalizaciones deverbales es esencial para el PLN (Subsección 1.1.2). No obstante, primero empezamos definiendo qué entendemos por nominalización deverbal y en qué tipo de nominalizaciones deverbales se centra este trabajo (Subsección 1.1.1).

1.1.1. ¿Qué nominalizaciones deverbales estudiamos?

El objeto de estudio de esta tesis son las nominalizaciones deverbales del español, es decir, sustantivos que mantienen una relación morfológica y/o semántica con el verbo del cual asumimos que heredan su estructura argumental. Proponemos una definición amplia de nominalización deverbal porque entre aquellas que nos interesan estudiar incluimos los sustantivos que realmente se derivan de verbos ('coronar > coronación'; 'lanzar > lanzamiento'; 'amenazar > amenaza') y algunos nombres a los que hemos llamado *cousin* siguiendo la terminología utilizada en NomBank (Meyers, 2007). Los sustantivos *cousin* no se derivan de verbos pero pueden mantener una relación morfológica o semántica con ellos: si la relación es morfológica, la derivación es siempre del sustantivo *cousin* al verbo correspondiente ('revolución > revolucionar'). Si no existe relación morfológica alguna, consideramos también como *cousin* aquellos sustantivos que mantienen una relación semántica con un verbo. Por ejemplo, 'victoria' se considera el sustantivo *cousin* de 'vencer' y 'éxito' la nominalización de la construcción con verbo soporte 'tener éxito'. Tratamos estos nombres en el grupo de los deverbales porque creemos que se les puede atribuir las mismas propiedades semánticas que los sustantivos deverbales morfológicamente derivados: pueden denotar eventos o resultados, pueden tener complementos que se interpreten como argumentos, pueden formar parte de construcciones lexicalizadas, etc. Por lo tanto, la noción de nominalización deverbal empleada en este trabajo es básicamente semántica, consideramos que una nominalización se relaciona semánticamente con un verbo

Nominalización deverbal

Nominalización *cousin*

independientemente de que derive morfológicamente o no de él.

Nominalización agentiva

Entre las diferentes nominalizaciones deverbales, nos centramos en aquellas que presentan la ambigüedad denotativa entre evento (o proceso) y resultado² (Grimshaw, 1990; Picallo, 1999; Pustejovsky, 1995; Badia, 2002), mientras que dejamos para un trabajo futuro las nominalizaciones deverbales agentivas (‘constructor’). Existen dos motivos por los cuales no incluimos las nominalizaciones deverbales agentivas en nuestro trabajo. Por un lado, no presentan la ambigüedad denotativa que estamos interesados en estudiar y, por otro, dado que se forman mediante una gama bien delimitada de sufijos (-ante, -ero, -or) son fácilmente identificables como nominalizaciones que tienen el argumento agente incorporado y para las que la anotación de la estructura argumental entrañaría, en principio, menos dificultades ya que el agente se expresa en la misma nominalización y el argumento con más probabilidad para ser explicitado, entonces, es el argumento interno, paciente o tema. En el ejemplo (3), la nominalización agentiva ‘cantante’ tiene el argumento agente incorporado (el que canta) por lo que el SP ‘de la última canción del disco’ se analiza como el argumento paciente de la nominalización.

(3) Me gusta la voz [del **cantante** de la última canción del disco]SN.

Nominalización
evento vs. resultado

En cuanto a la distinción denotativa, entendemos por nominalización de evento aquella nominalización que denota una acción o un proceso de la misma forma que un verbo los denota. En otras palabras, las nominalizaciones eventivas, de la misma manera que los verbos correspondientes, tienen la propiedad aspectual de la dinamicidad (4). En cambio las nominalizaciones resultativas se caracterizan por denotar estados (6) o el objeto, concreto o abstracto, resultante de una acción (5). Ambos tipos de nominalizaciones resultativas (estados y objetos) carecen de la propiedad aspectual de la dinamicidad.

(4) El proyecto americano consiste en [la **adaptación**_{<evento>} de la novela *Paper Boy*]SN.

(5) [Esta **adaptación**_{<resultado>} cinematográfica]SN ha recibido buenas críticas.

(6) Reforzó [la **tendencia**_{<resultado>} al alza del Euro de los últimos días]SN.

En el ejemplo (4), el sustantivo ‘adaptación’ denota un evento porque en él se expresa una acción de la misma manera que un verbo la podría expresar (‘El proyecto americano consiste en adaptar la novela *Paper Boy*’). La interpretación eventiva se caracteriza como dinámica porque implica un cambio de estado: desde el estado

²La diferenciación de evento vs. resultado es la distinción denotativa más extendida entre los autores, si bien, como veremos en el Capítulo 2, existen distintas tipologías de nominalizaciones deverbales.

de ‘no estar adaptado’ al estado de ‘de estar adaptado’. En cambio, en el ejemplo (5) la misma nominalización se entiende como un resultado porque denota un objeto específico que es el producto de la acción de adaptar una obra a una película. En el ejemplo (6), la interpretación resultativa se origina porque el verbo base de ‘tendencia’, ‘tender’ denota un estado, por lo que el sustantivo hereda la propiedad aspectual de la estatividad (no dinamicidad) que no implica ningún cambio de estado.

1.1.2. La importancia para el PLN de las nominalizaciones deverbales

Desde un punto de vista lingüístico, dos son los temas claves que atañen a las nominalizaciones deverbales: su tipo denotativo (la distinción evento *vs.* resultado mencionada) y su estructura argumental, y ambas cuestiones son importantes para el PLN. El tipo denotativo se refiere al tipo de interpretación semántica de la nominalización deverbal, ya se interprete como un evento o un resultado, según la definición ofrecida. La estructura argumental se refiere al conjunto ordenado de argumentos de la nominalización que completan su significado.

Resolución de la
correferencia

En cuanto al tipo denotativo de las nominalizaciones deverbales, tener detectada esta diferencia semántica puede ser útil en un sistema automático de resolución de la correferencia. Esta tarea consiste en identificar en un texto qué SNs o menciones se refieren a la misma entidad (Recasens, 2010). Concretamente, conocer la denotación podría ayudar a identificar tipos de correferencia. Por ejemplo, si una nominalización deverbal tiene un antecedente y el tipo denotativo es eventivo, se puede establecer una relación correferencial de identidad entre ellos (7). En cambio, si la nominalización es resultativa, se establecería una relación correferencial puente (*bridging coreference*) (8) (Clark, 1975; Recasens et al., 2007)³.

- (7) En Francia los precios **cayeron** un 0,1 % en septiembre. [**La caída**_{<evento>}]SN ha provocado que la inflación quedara en el 2,2 %.
- (8) La imprenta **se inventó** en el año 1.440. [**El invento**_{<resultado>}]SN permitió difundir las ideas y conocimientos con eficacia.

Detección de paráfrasis

Reconocer esta diferencia semántica también puede resultar muy útil para la tarea de detección de paráfrasis (Androutsopoulos and Malakasiotis, 2010; Madnani and Dorr, 2010; Vila et al., 2011), que consiste en reconocer si dos expresiones del lenguaje concretas constituyen una paráfrasis o no. Las nominalizaciones eventivas, pero no las resultativas, pueden ser paráfrasis de cláusulas verbales,

³Los criterios usados para resolver estos dos tipos de correferencia son distintos por lo que la distinción es importante computacionalmente.

por lo que saber la denotación puede ayudar a detectar paráfrasis. Por ejemplo, la oración en (9) se considera paráfrasis del SN en (10), un SN cuyo núcleo es una nominalización eventiva.

(9) **Se ha ampliado** el capital de la empresa en un 20 %.

(10) [**La ampliación**_{<evento>} del capital de la empresa en un 20 %]_{SN}.

Sin embargo, si la nominalización tiene una interpretación resultativa como en (11) –‘traducciones’ se refiere al objeto concreto, es decir, al libro traducido–, es imposible tener una paráfrasis con una cláusula verbal. Esto se explica porque las nominalizaciones deverbales resultativas denotan objetos, mientras que en los verbos es imposible porque denotan acciones. En este sentido, las nominalizaciones deverbales resultativas solo pueden ser paráfrasis de otros SNs que denoten objetos (12).

(11) Se han vendido [**muchas traducciones**_{<resultado>} de su último título]_{SN}.

(12) Se han vendido [**muchos libros traducidos** de su último título]_{SN}.

WSD

Además, si anotamos esta diferencia semántica en el corpus AnCora-Es, este corpus podría ser utilizado también como corpus de entrenamiento y evaluación de sistemas de WSD, cuyo objetivo es identificar qué sentido de una palabra es correcto en un contexto determinado cuando esa palabra se caracteriza por ser polisémica. En este caso, serviría para entrenar sistemas que diferenciaran entre los diferentes sentidos de las nominalizaciones (sentidos eventivos vs. sentidos resultativos).

Aplicaciones: EI, RI, BdR

Respecto a la estructura argumental de las nominalizaciones deverbales, en este trabajo partimos de la hipótesis de que estas heredan la estructura argumental de los verbos base correspondientes y, al igual que estos, expresan relaciones semánticas de tipo argumental y temático (agente, paciente, causa, etc.). Por lo tanto, de la misma manera que en los verbos, tener identificadas dichas relaciones puede ser muy útil para cualquier tarea o aplicación de PLN, especialmente, Extracción y Recuperación de Información y Sistemas de Búsqueda de Respuestas. Por ejemplo, ante la pregunta ‘¿Quién inventó la bombilla?’ podemos encontrar la respuesta ‘Edison inventó la bombilla’ o ‘La invención de la bombilla por Edison’. Por lo tanto, si queremos detectar ambas respuestas es necesario que tengamos representados semánticamente tanto oraciones como SNs con núcleo de sustantivo verbal.

1.2. Objetivos del trabajo

El objetivo principal de esta investigación es el estudio lingüístico de las nominalizaciones deverbales descritas en la Subsección 1.1.1. Concretamente nos centramos en la estructura argumental y en el tipo denotativo de las nominalizaciones deverbales del español. Este estudio nos permite representarlas semánticamente con el objetivo de que puedan ser procesadas computacionalmente. Este objetivo principal se desglosa en cinco objetivos concretos resumidos a continuación:

1. Estudio lingüístico basado en corpus de las nominalizaciones deverbales: ambigüedad denotativa y estructura argumental.
2. Desarrollo de un sistema automático que permita establecer la distinción de sustantivos deverbales según su denotación (ADN-Classifer).
3. Desarrollo de un sistema automático que permita anotar la estructura argumental de las nominalizaciones deverbales (RHN).
4. Enriquecimiento del corpus AnCora-Es con la anotación (denotación y estructura argumental) de las nominalizaciones deverbales.
5. Creación de un léxico, AnCora-Nom, como recurso lingüístico que representa las nominalizaciones deverbales en español.

El primero de los objetivos es realizar un estudio empírico sobre las nominalizaciones deverbales del español con el fin de observar su comportamiento sintáctico-semántico. Para llevarlo a cabo, se ha utilizado el corpus AnCora-Es (Taulé et al., 2008) del que obtenemos la casuística a analizar. Nótese que el corpus AnCora-Es es tanto la fuente de información de la que partimos, como el corpus que se enriquecerá con la anotación de los sustantivos deverbales (objetivo 3). Para realizar el análisis lingüístico es necesario, primero, revisar las diferentes propuestas teóricas respecto a la denotación y la estructura argumental, con el propósito de elaborar una primera lista de criterios o aspectos a tener en cuenta que servirá de base para el análisis de los datos. Después, se lleva a cabo un análisis empírico de los datos en base a la propuesta teórica inicial. Este análisis se valdrá, además de la interpretación del lingüista, de técnicas de aprendizaje automático que nos ayuden a refrendar empíricamente nuestras hipótesis e intuiciones y a evaluar cuantitativamente la importancia de los diferentes factores que consideremos.

Este análisis lingüístico nos proporcionará los rasgos que caracterizan a las nominalizaciones deverbales del español respecto a los dos aspectos que estudiamos, denotación y estructura argumental. Así, por ejemplo, tras el análisis lingüístico, detectamos los rasgos lingüísticos que mejor distinguen las nominalizaciones

eventivas y resultativas y observamos patrones de realización sintáctica de los argumentos de los sustantivos deverbales. Estas observaciones las implementamos en dos sistemas automáticos que servirán para anotar la denotación (ADN) y la estructura argumental (RHN) de las nominalizaciones deverbales en el corpus AnCora-Es. La evaluación de dichos sistemas, por tanto, también supondrá la evaluación de las hipótesis lingüísticas de partida. Sin embargo, el objetivo de la creación de estos sistemas no es solo que se utilicen para la anotación de estos dos tipos de información en el corpus AnCora-Es, sino que son sistemas desarrollados con la intención de que constituyan herramientas para la comunidad científica, es decir, herramientas que puedan ser utilizadas con otros corpus y en otros contextos.

Más allá de los tres objetivos descritos hasta el momento (estudio lingüístico y los dos sistemas automáticos) esta tesis tiene también como objetivo la creación de recursos lingüísticos para el español, una lengua en la que las nominalizaciones deverbales aún no están ampliamente estudiadas y que carecen de recursos léxicos que soporten su estudio. Así, por lo tanto, tenemos como objetivo anotar semánticamente (denotación y estructura argumental) los SNs de núcleo de verbal del corpus AnCora-Es. A partir de esta anotación, y sobre todo de su validación manual, obtenemos la evaluación de los sistemas automáticos desarrollados, de las hipótesis lingüísticas subyacentes y nuevas observaciones lingüísticas. Una vez anotado y validado el corpus AnCora-Es, un último objetivo ha sido elaborar un léxico nominal, AnCora-Nom, en el que queda representada toda la información que atañe a las nominalizaciones deverbales.

1.3. Procedimiento

En la Figura 1.1 se muestra el procedimiento seguido para la elaboración de esta tesis, que consta de tres grandes etapas identificadas por la gradación del sombreado.

Etapas 1

El primer paso fue estudiar y analizar qué se había dicho y hecho sobre las nominalizaciones deverbales con anterioridad. Una de las cuestiones más tratadas desde un punto de vista teórico era la distinción de evento y resultado que ocupa a estas nominalizaciones, y que en algunos autores iba ligada a la capacidad argumental de las mismas. Ante este hecho, decidimos en primer lugar llevar a cabo un estudio lingüístico basado en corpus, que nos permitiese analizar empíricamente algunas de las afirmaciones sostenidas por dichos autores. Nos centramos básicamente en evaluar, por un lado, qué criterios propuestos en la bibliografía para esta distinción son válidos para el español y en detectar nuevos posibles criterios

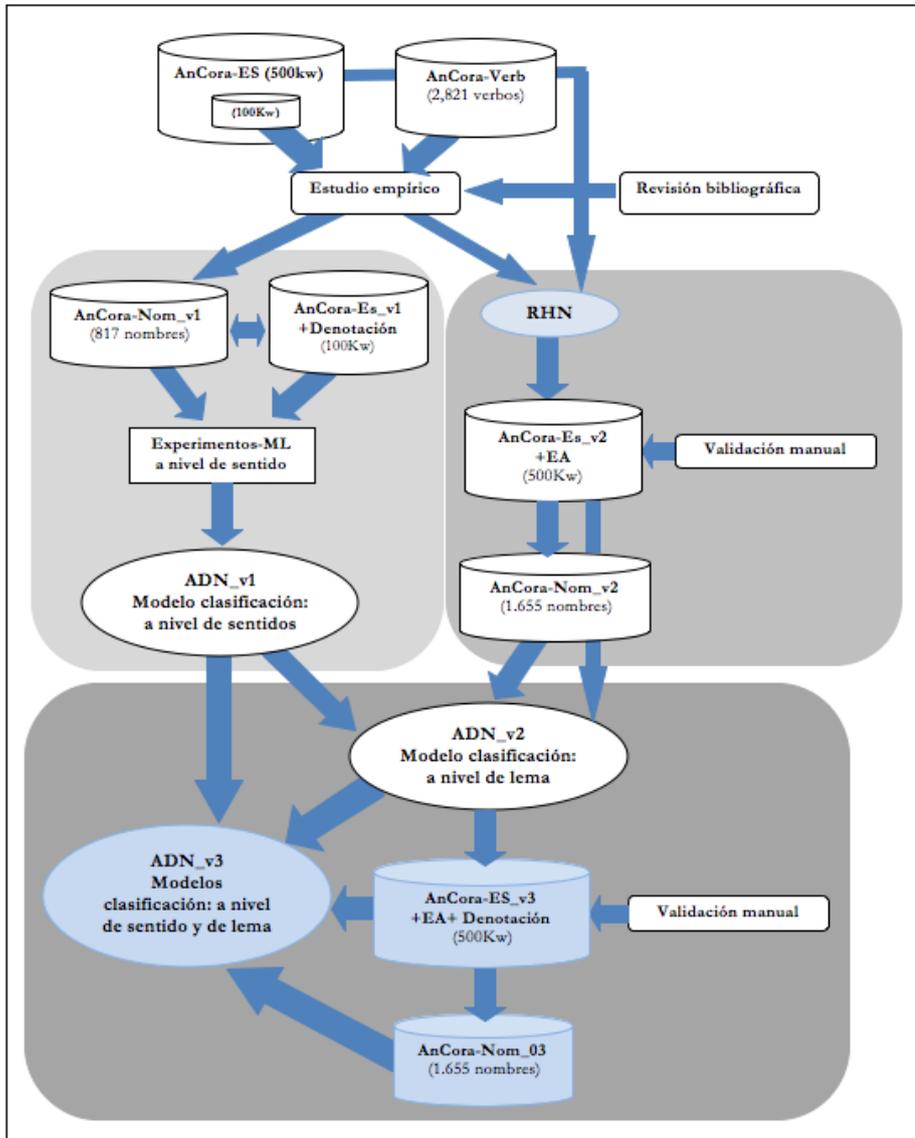


Figura 1.1: Esquema del procedimiento utilizado en el desarrollo de la tesis

y, por el otro lado, estudiamos las diferentes estructuras sintácticas mediante las cuales se realizan los argumentos de las nominalizaciones deverbales (Peris and Taulé, 2009). Este estudio empírico se realizó sobre un subconjunto de 100.000 palabras del corpus AnCora-Es que corresponden al corpus originario 3LB (Civit and Martí, 2004), y dio lugar a un léxico construido manualmente de 817 entradas nominales (AnCora-Nom-v1), correspondientes a los 817 lemas diferentes de nominalizaciones deverbales que se encuentran en este subconjunto del corpus. Este

primer léxico, en el que se representa la información sintáctico-semántica asociada a las nominalizaciones, nos permitió anotar manualmente las ocurrencias correspondientes (un total de 3.077) de este subconjunto del corpus (AnCora-Es-v1). En lo que respecta a la distinción denotativa, dado que la dificultad en establecer la diferencia entre evento y resultado resultaba a veces muy complicada, quisimos evaluar la modelización de esta distinción subyacente en los recursos creados (AnCora-Nom-v1 y AnCora-Es-v1, en la Figura 1.1), es decir, se evaluaron los atributos relacionados con dicha distinción. Para ello, aplicamos técnicas de aprendizaje automático (*Machine Learning*, ML) sobre el léxico AnCora-Nom-v1. El modelo de clasificación resultante está basado en distinciones de sentido, esto es, la extracción de los rasgos se realizó a nivel de sentido y las instancias para el aprendizaje se corresponden con los sentidos del léxico AnCora-Nom-v1. Nos referimos a este primer modelo de clasificación como ADN-v1 a nivel de sentido. Se realizaron una serie de experimentos con técnicas de ML utilizando la plataforma Weka (Witten and Frank, 2005), tanto para el proceso de aprendizaje del clasificador como para el de clasificación, con el objetivo de evaluar la consistencia de los datos anotados en este primer léxico, de analizar la relevancia de los atributos utilizados en la representación de las nominalizaciones deverbales y de inferir nuevos atributos para la representación de esta distinción (Peris et al., 2009). De esta manera, sentamos las bases para la construcción de un clasificador automático de nominalizaciones deverbales según su denotación.

Etapas 2

Tras este estudio empírico, nuestro objetivo consistía en la anotación de la estructura argumental y del tipo denotativo de todas las ocurrencias de nominalizaciones deverbales en AnCora-Es. Sin embargo, puesto que el número de ocurrencias en el corpus es elevado (alrededor de 24.000), se decidió que la anotación debía realizarse de forma automática con un proceso posterior de validación manual. Uno de los resultados del análisis empírico es que resulta casi imposible establecer la diferencia denotativa sin tener en cuenta la información de la estructura argumental, por lo que la anotación de los argumentos semánticos se confirmó como el siguiente paso en este proceso de investigación. De esta manera, a partir de los patrones de realización sintáctica de los argumentos de los sustantivos deverbales observados, creamos un sistema automático basado en reglas –RHN (Peris and Taulé, 2011b)– que permitió la anotación automática de los argumentos de todas las ocurrencias de nominalizaciones deverbales del corpus AnCora-Es (23.431 nominalizaciones, en total), que se corresponden a 1.655 lemas diferentes (AnCora-Es-v2 en la Figura 1.1). Las RHN son reglas de proyección que parten principalmente de la información codificada en el léxico verbal, AnCora-Verb (Aparicio et al., 2008), y que se aplican en formato de lista de decisión. Tras este proceso automático, el corpus enriquecido con la estructura argumental de las no-

nominalizaciones fue validado manualmente, lo que permitió así la evaluación de las reglas de proyección diseñadas y de las observaciones lingüísticas subyacentes. A partir de esta anotación en el corpus se creó una nueva versión del léxico nominal, AnCora-Nom-v2 en la Figura 1.1, que contenía 1.655 entradas nominales, correspondientes a todas las nominalizaciones deverbales del corpus. AnCora-Nom-v2 tenía incorporada información acerca de la estructura argumental y otros atributos del léxico inicial excepto la denotación. Con este nuevo léxico nos concentramos en la tarea de construir el clasificador automático de nominalizaciones deverbales según su denotación.

Etapa 3

El objetivo de esta tercera etapa era desarrollar el clasificador ADN (Anotación de la Denotación en la Nominalizaciones deverbales), un sistema de clasificación automática de este tipo de sustantivos según su denotación. Con dicho objetivo, incrementamos la muestra de datos a partir de la cual el clasificador debía aprender. Por lo tanto, era necesario anotar la denotación en todas las ocurrencias de nominalizaciones deverbales de AnCora-Es. Dado que esto implicaba un notable aumento de las ocurrencias a anotar (23.431 ocurrencias en comparación con las 3.077 iniciales), se optó por que se realizara de manera automática. Para tal propósito se adaptó el modelo de clasificación ADN-v1 a nivel de sentido aprendido con anterioridad (en la Etapa 1) a un modelo de clasificación a nivel de lemas, ADN-v2, con el objetivo de que pudieran clasificarse automáticamente las ocurrencias de nominalizaciones del corpus AnCora-Es según su tipo denotativo. Este nuevo modelo parte de los siguientes recursos: 1) el léxico AnCora-Verb, para obtener los rasgos relacionados con los verbos correspondientes a las nominalizaciones; 2) el corpus AnCora-Es al completo (500.000 palabras), del que se obtienen distintos rasgos morfosintácticos y semánticos; y 3) del recién creado léxico AnCora-Nom-v2, del que obtiene, entre otras, la información sobre la estructura argumental de todos los lemas de las nominalizaciones del corpus. Sin embargo, dado que ADN-v2 trabaja a nivel de lema, prescinde de toda información que sea específica de un sentido determinado y solo tiene en cuenta la información compartida por todos los sentidos de un mismo lema, lógicamente esta granularidad menos fina tiene el coste de una caída en la precisión del clasificador ADN (Peris et al., 2010a). El nuevo modelo de clasificación a nivel de lema, ADN-v2, se utilizó para la anotación automática del tipo denotativo en el corpus AnCora-Es. Con el objetivo de evaluar el rendimiento de este modelo, el corpus fue manualmente validado (Peris et al., 2010b), dando lugar a una nueva y definitiva versión de AnCora-Es (-v3) en la Figura 1.1. A partir de este corpus manualmente validado se generó la versión final del léxico AnCora-Nom (-v3) en la Figura 1.1, que incluye también información sobre el tipo denotativo para todas las entradas léxicas (Peris and Taulé, 2011a).

Finalmente, para construir la versión última del clasificador ADN (ADN-v3 en la Figura 1.1), se realizaron una serie de experimentos con el objetivo de construir nuevos modelos de clasificación (a nivel de sentido y a nivel de lema) a partir de los recursos recién creados, es decir, AnCora-Nom-v3 y AnCora-Es-v3, unos modelos que aprenden con un mayor número de instancias y con recursos totalmente validados. Además también se replicaron los experimentos a nivel de sentido y lema con el subconjunto de AnCora-Es-v3 de 100.000 palabras y el subconjunto de 817 entradas léxicas de AnCora-Nom-v3 de la primera etapa (Peris et al., 2012). Para la evaluación de todos estos nuevos modelos desarrollados, basados en sentidos y basados en lemas, se ha utilizado la validación cruzada con 10 particiones aleatorias, *ten fold cross-validation* a partir de AnCora-Nom-v3 y AnCora-Es-v3. Estos modelos dan lugar a la versión final del clasificador ADN (ADN-v3). En la Figura 1.1 se han sombreado en azul las dos herramientas –RHN y ADN– desarrolladas en el marco de este trabajo para el tratamiento computacional de las nominalizaciones de verbales, la primera centrada en la estructura argumental y la segunda en la denotación. También encontramos en azul los dos recursos derivados de este trabajo de investigación, el léxico nominal AnCora-Nom-v3 y el corpus anotado AnCora-Es-v3, que pueden ser utilizados tanto como fuente de consultas lingüísticas así como corpus de aprendizaje para sistemas computacionales de SRL (estructura argumental) o WSD (tipo denotativo).

1.4. Contribuciones

Las contribuciones de esta tesis, que se presentan a continuación, están estrechamente relacionadas con los objetivos propuestos.

- Conjunto de criterios lingüísticos que permiten establecer la distinción de evento y resultado en español. Estos criterios se han obtenido a partir del estudio empírico sobre el 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, son los llamados selectores. Además, a partir de la observación de las reglas creadas por el clasificador, se ha identificado algún criterio más para establecer la distinción entre evento y resultado de las nominalizaciones.
- Estudio lingüístico de la estructura argumental de las nominalizaciones de verbales, 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 que se aplican en un formato de lista de decisión, que ha permitido anotar la estructura argumental de las nominalizaciones deverbales del corpus AnCora-Es.
- Guía de anotación para la validación manual del corpus referente a la estructura argumental de las nominalizaciones deverbales.
- Guía de anotación para la validación manual del corpus referente al tipo denotativo de las nominalizaciones deverbales.
- Adaptación de la herramienta AnCora-Pipe para la anotación de las nominalizaciones deverbales.
- Enriquecimiento del corpus AnCora-Es con sendas validaciones manuales 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.
- Primera guía para la anotación de los argumentos implícitos en el corpus AnCora-Es.

1.5. Estructura 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 verbal, la importancia de su estudio y ofrece una panorámica de los trabajos realizados, tanto desde el punto de vista lingüístico como computacional, sobre este tipo de construcción. La segunda parte centra su atención en la estructura argumental de las nominalizaciones deverbales, tanto el estudio empírico realizado sobre este aspecto como el sistema automático desarrollado (RHN) para la anotación de dicha información en el corpus. La tercera parte trata la distinción denotativa entre evento y resultado que afecta a las nominalizaciones deverbales, tanto el estudio empírico realizado sobre este aspecto, como el sistema de clasificación automático desarrollado (ADN) para la anotación de dicha información en el corpus y los

experimentos desarrollados con este clasificador. Finalmente, 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 y el léxico derivado AnCora-Nom. Estas cuatro partes comprenden los siguientes capítulos.

Parte I: Antecedentes

- El **Capítulo 1** da cuenta de la importancia de estudiar las nominalizaciones deverbales, tanto por su riqueza semántica como por la utilidad de su tratamiento computacional para tareas de PLN. Además, se delimita el objeto de estudio, es decir, se define qué entendemos por nominalización verbal. En este capítulo introductorio también se definen los objetivos del trabajo y el procedimiento seguido para lograrlos. A continuación se detallan las contribuciones de esta investigación y, finalmente, se cierra el capítulo con la estructura de la tesis.
- El **Capítulo 2** recoge una revisión bibliográfica de los trabajos sobre las nominalizaciones deverbales. Abarcamos tanto las aproximaciones teóricas como las computacionales, aunque podemos avanzar que mientras que los trabajos teóricos se centran tanto en la distinción del tipo denotativo como en la estructura argumental, en los trabajos computacionales cobra más relevancia este segundo aspecto, si bien hay trabajos computacionales que también trabajan sobre la diferenciación denotativa.

Parte II: Estructura argumental

- El **Capítulo 3** presenta el estudio empírico basado en corpus sobre los patrones sintácticos de realización de los argumentos de las nominalizaciones deverbales. En este capítulo también se detalla el esquema de anotación utilizado para la anotación de la estructura argumental, que será el mismo que se utiliza en la anotación automática y la posterior validación manual.
- El **Capítulo 4** explica cómo se ha llevado a cabo la anotación automática de los argumentos internos al SN en el corpus AnCora-Es: las reglas heurísticas (RHN) propuestas a partir de las observaciones obtenidas en el estudio empírico, su evaluación y, por consiguiente, el enriquecimiento del corpus AnCora-Es con la anotación de la estructura argumental de las nominalizaciones deverbales.

Parte III: Denotación

- El **Capítulo 5** presenta el estudio empírico basado en corpus sobre la distinción denotativa entre evento y resultado. Incluye el análisis lingüístico que determina qué criterios de la bibliografía son válidos para el español y ofrece nuevos criterios lingüísticos que ayudan a establecer dicha distinción. También incluye el análisis computacional realizado para evaluar con métodos de ML los resultados obtenidos.
- El **Capítulo 6** describe el clasificador ADN, un sistema de clasificación automática de denotaciones deverbales según su denotación, y los recursos que se han utilizado para el desarrollo de este clasificador.
- El **Capítulo 7** explica los distintos experimentos aplicados a ADN, en los que se han desarrollado distintos modelos de clasificación, tanto a nivel de sentido como de lema, y la evaluación de dichos modelos.

Parte IV: Recursos

- El **Capítulo 8** detalla los dos procesos de validación manual llevados a cabo, los criterios de anotación tanto para la estructura argumental como para la denotación y las pruebas de acuerdo entre anotadores que garantizan la fiabilidad de la validación manual. Además se describe la extensión de la herramienta AnCora-Pipe (Bertran et al., 2008) realizada para la anotación de los SNs en el corpus AnCora-Es.
- El **Capítulo 9** explica el proceso de inducción del léxico AnCora-Nom a partir de la información anotada en el corpus AnCora-Es y detalla la información contenida en dicho léxico.

Finalmente, en el **Capítulo 10** se recogen las conclusiones globales de este trabajo, las aportaciones del mismo y las líneas de trabajo futuro, entre las que cobra una especial relevancia el estudio de los argumentos implícitos de las nominalizaciones deverbales, una línea en la que ya se ha empezado a trabajar.

CAPÍTULO 2

NOMINALIZACIONES DEVERBALES: ESTADO DE LA CUESTIÓN

En este capítulo se presenta una revisión bibliográfica de los trabajos que tratan las nominalizaciones deverbales, tanto desde un punto de vista lingüístico como computacional. Desde una perspectiva lingüística (Sección 2.1), distinguimos entre aquellos que definen las nominalizaciones deverbales en función básicamente de su tipo denotativo (Sección 2.1.1) y aquellos que se centran principalmente en su estructura argumental (Sección 2.1.2). Desde una perspectiva computacional (Sección 2.2) revisamos, por un lado, los recursos existentes para distintas lenguas que representan las nominalizaciones deverbales (Sección 2.2.1) y, por el otro, los sistemas automáticos creados para su tratamiento computacional (Sección 2.2.2).

2.1. Aproximaciones lingüísticas

En este apartado abordamos el tema de la nominalización deverbal desde diferentes marcos teóricos. Hemos tenido en cuenta estudios sobre la nominalización desde la Gramática Generativa, la Gramática Léxico-Funcional, la Teoría Sentido-Texto, el Lexicón Generativo, la Gramática del Rol y la Referencia, la Gramática de Estructura Sintagmática regida por el Núcleo, así como desde un punto de vista esencialmente descriptivo-gramatical. Estas aproximaciones coinciden en el hecho de que clasifican las nominalizaciones deverbales según su denotación, aunque algunas se centran más específicamente en la estructura argumental. A pesar de que ambos aspectos están relacionados y que la mayoría de los autores que tratan las nominalizaciones deverbales contemplan en mayor o menor medida ambos aspectos, se ha decidido, por razones expositivas, estructurar esta sección en dos

subapartados: en el primero (Subsección 2.1.1) presentamos aquellos trabajos que se centran en la denotación y en el segundo (Subsección 2.1.2) aquellos que lo hacen en la estructura argumental. Respecto a la denotación, se presentan las distintas propuestas tipológicas de nominalizaciones deverbales según su denotación; se atiende a la polémica entre algunos autores por considerar las distintas denotaciones de las nominalizaciones como sentidos de una misma unidad léxica o bien como unidades léxicas diferentes; y, finalmente, se recogen los diferentes criterios propuestos para la distinción de las nominalizaciones según su denotación. En cuanto a la estructura argumental, la revisión bibliográfica se centra en las diversas propuestas de representación de esta en las nominalizaciones deverbales.

2.1.1. Nominalizaciones deverbales y denotación

La mayor parte de los autores que tienen en cuenta la denotación en la caracterización de las nominalizaciones deverbales distinguen básicamente entre la denotación de evento (1) y la de resultado (2), aunque la terminología utilizada no es siempre la misma. Como vimos en el Capítulo 1, las nominalizaciones de evento están caracterizadas por poseer la propiedad de la dinamicidad y denotan acciones, de la misma manera que los predicados verbales. En el ejemplo (1), el predicado nominal ‘combinación’ denota una acción al igual que la paráfrasis verbal equivalente: ‘lo que condujo a que se combinaran para...’. Las nominalizaciones de resultado, en cambio, carecen de dicha dinamicidad, por lo que bien denotan un estado o el resultado de la acción expresada por el verbo base correspondiente. En el ejemplo (2), el sustantivo deverbale ‘combinación’ denota en este caso el resultado de la acción de ‘combinar’. A pesar de que la clasificación de las nominalizaciones deverbales en estas dos denotaciones es la más generalizada, existen propuestas de clasificación más finas que se presentarán a lo largo de esta sección.

- (1) Lo que condujo a [[su]_{POSS} **combinación** [para formar el complejo n-molecular dador aceptor]_{SP}]_{SN}.
- (2) De [dicha **combinación**]_{SN} nace una criatura con características propias.

Entre los autores que clasifican las nominalizaciones deverbales en eventos y resultados, existen dos cuestiones en las que los distintos autores discrepan: a) la representación léxica de las dos denotaciones y b) la capacidad argumental de las nominalizaciones deverbales.

En cuanto a la primera polémica, algunos consideran estas dos denotaciones como dos unidades léxicas diferentes (Grimshaw, 1990; Alexiadou, 2001; Picallo, 1999) mientras que otros sostienen que ambas denotaciones son sentidos de una misma unidad léxica (Pustejovsky, 1995; Mel’cuk et al., 1984; Alonso, 2004).

2. NOMINALIZACIONES DEVERBALES: ESTADO DE LA CUESTIÓN

Para diferenciar estas dos denotaciones a nivel léxico, los representantes de la primera posición son los que con más profusión proponen criterios morfosintácticos y semánticos para diferenciarlas.

La segunda polémica tiene que ver con la capacidad argumental, negada por algunos autores a los sustantivos resultativos. Grimshaw (1990) y Zubizarreta (1987) consideran que la capacidad argumental se circunscribe a los sustantivos eventivos, mientras que autores como Alexiadou (2001), Picallo (1999), Pustejovsky (1995) y Mel'cuk et al. (1984) mantienen que tanto las nominalizaciones de evento como de resultado pueden legitimar argumentos.

A continuación presentamos brevemente las propuestas de los principales autores en referencia a las polémicas presentadas.

Desde la Gramática Generativa (*Generative Grammar*, GG en adelante), una de las primeras propuestas de clasificación es la de Zubizarreta (1987). Esta autora clasifica las nominalizaciones deverbales del inglés en cuatro tipos en función de si tienen estructura léxico-semántica (S-R)¹, de su denotación y de si son o no contables: 1) las nominalizaciones derivadas de verbos estativos no tendrían estructura S-R (*love* 'amor', *hatred* 'odio', *fear* 'miedo'); 2) las que denotan un proceso o un resultado y de las que se asume que tienen una estructura S-R opcional (*description* 'descripción', *translation* 'traducción', *proposal* 'propuesta'); 3) las que tan solo pueden denotar un proceso, que además son contables [+contable] y cuya estructura S-R sería opcional (*assassination* 'asesinato', *capture* 'captura', *coronation* 'coronación'); y finalmente, 4) las que solo denotan un proceso pero que, contrariamente a las anteriores, no son contables [-contable] y tienen una estructura S-R obligatoria (*destruction* 'destrucción', *recognition* 'reconocimiento', *memorization* 'memorización'). Dado que la estructura léxico-semántica (S-R) es la que codifica las restricciones selectivas de los ítems léxicos, podríamos inferir que aquellas nominalizaciones que no tengan S-R no legitimarían argumentos; las que tengan un S-R opcional pueden legitimarlos de manera optativa y, finalmente, las que presenten un S-R obligatorio, que se vincula con el rasgo [-contable], siempre tienen argumentos. Cabe destacar, sin embargo, que esta autora no aclara si proceso y resultado son las dos únicas clases de nominalización de verbal. ¿Qué ocurre con las nominalizaciones derivadas de verbos estativos, constituyen una clase aparte?

Generative Grammar

Zubizarreta, 1987

Grimshaw (1990), también desde el marco de la GG, establece por primera vez de manera explícita la relación entre la capacidad argumental de las nominalizaciones deverbales y su denotación. Esta autora distingue para el inglés tres tipos de sustantivos en función de su denotación: 1) aquellos que denotan un evento

Grimshaw, 1990

¹Esta estructura léxico-semántica, resumida como S-R, es, a nuestro entender, muy similar a la noción de estructura argumental puesto que es la estructura en la que se codifican las restricciones argumentales de los ítems léxicos.

complejo² (*examination*, ‘revisión’); 2) los que denotan un evento simple³ (*trip*, ‘viaje’); y 3) los que denotan el resultado⁴ de una acción (*exam*, ‘examen’)⁵. Para ella, esta diferencia denotativa está estrechamente vinculada con la capacidad de selección de argumentos: tan solo los sustantivos eventivos complejos legitiman una estructura argumental y, por consiguiente, solo ellos seleccionan argumentos. Las otras dos clases carecen de estructura argumental y, en consecuencia, no seleccionan argumentos, aunque sí tienen estructura léxico-conceptual y, por lo tanto, pueden tener participantes, que es un concepto similar al de argumento pero no legitimado a nivel sintáctico-semántico. Las nominalizaciones deverbales solo pueden ser del primer y tercer tipo de sustantivos propuestos por Grimshaw (los eventos simples son sustantivos que no derivan de verbos pero que denotan un evento), por lo tanto, podemos decir que ella distingue entre nominalizaciones de evento complejo y nominalizaciones de resultado. De hecho, la autora propone una serie de criterios lingüísticos que están orientados a justificar esta doble distinción (denotativa y de capacidad argumental); una contribución en profundidad que hasta ese momento no se había producido. De la importancia de este trabajo da cuenta el hecho de que todos los investigadores que trabajan sobre nominalizaciones deverbales se refieren a este estudio bien para sostenerlo, bien para cuestionarlo.

- Borer, 1997** En el paradigma de la GG, en la misma línea que Grimshaw se encuentra el trabajo de Borer (1997), también para el inglés, que argumenta que las propiedades de las nominalizaciones de proceso derivadas (eventos complejos en la terminología de Grimshaw) deben estar relacionadas con las propiedades del verbo del que derivan. Borer postula que la estructura argumental aparece en nominalizaciones de proceso derivadas puesto que en ellas está presente un sintagma verbal (SV) totalmente proyectado en la estructura sintáctica de la nominalización y es este SV proyectado el que se encarga de asignar papeles temáticos a los argumentos. Según esta autora, lo que diferencia a las nominalizaciones derivadas resultativas de las de proceso es que en las primeras no hay proyección del SV y, por lo tanto, tampoco hay estructura argumental. Para el español, mantiene una tesis similar Gràcia i Solé (1995) quien argumenta que las nominalizaciones deverbales eventivas heredan la estructura argumental del verbo base correspondiente y, por el contrario, las nominalizaciones deverbales resultativas bloquean esta herencia.
- Gràcia i Solé, 1995**
- Demonte, 1989** También sobre el español es el trabajo de Demonte (1989) quien apoya el punto de vista de Grimshaw (1990) al considerar a los complementos de nombres resultativos como meros participantes y no argumentos. Esta misma hipótesis también

²*Complex event.*

³*Simple event.*

⁴*Result.*

⁵La diferencia entre evento simple (*simple event*) y evento complejo (*complex event*) radica en que solo los segundos son sustantivos derivados de verbos.

es secundada por Martí i Girbau (2002) para el catalán.

Martí i Girbau, 2002

Desde nuestro punto de vista, el problema principal de asociar las nominalizaciones eventivas con la capacidad de tener estructura argumental y negar esta posibilidad a las resultativas, es que las razones argumentadas para ello (la diferencia entre participante y argumento, paralela a la distinción entre estructura léxico-conceptual y estructura argumental, y la proyección de un SV en las primeras y no en las segundas) no pueden ser comprobadas de manera empírica, es decir, no se puede llevar a cabo un estudio basado en corpus que permita verificar la existencia de un SV en las nominalizaciones o diferencie si los complementos de las nominalizaciones son argumentos o participantes. Por este motivo nos resulta cuestionable la afirmación de que las nominalizaciones resultativas no tienen estructura argumental.

Sin embargo, no siempre desde la GG se excluye la posibilidad de que las nominalizaciones resultativas tengan estructura argumental. Para Alexiadou (2001) la diferencia entre sustantivos resultativos y de proceso (eventos complejos en términos de Grimshaw y eventos en términos de autores como Picallo (1999)) no radica en la estructura argumental, sino en la presencia de proyecciones de Voz y Aspecto en la estructura funcional de los eventivos, marcas típicas de los verbos. Esta afirmación está secundada por un estudio del comportamiento de las nominalizaciones en diferentes lenguas. La autora apunta que en muchas lenguas que tienen morfemas que definen diferencias de Voz (como ocurre en griego, maorí, turco y coreano, entre otras) y Aspecto (como en las lenguas eslavas) dentro del dominio verbal, estos mismos morfemas se utilizan en construcciones con nominalizaciones de proceso. Este comportamiento morfosintáctico difiere del de las nominalizaciones resultativas, las cuales (en las mismas lenguas) no pueden aparecer ni con adverbios aspectuales y de manera, ni con morfemas de Voz y Aspecto. En su estudio, esto constituye una confirmación de la existencia de las proyecciones de Voz y Aspecto en las nominalizaciones de proceso. Así, por tanto, no siendo la capacidad argumental la diferencia entre los dos tipos denotativos, esta autora afirma que ambos tipos de nominalizaciones pueden tener argumentos: *“Given that there is no lexical difference between verbs and process nouns, and between result and process nouns, apart from the functional domain, all can take arguments”* (Alexiadou, 2001, p. 69). Esto constituye una diferencia fundamental respecto al trabajo de Grimshaw.

Alexiadou, 2001

Picallo (1999) centra su estudio en las nominalizaciones deverbales del español por lo que es especialmente interesante para nuestro trabajo. Esta autora mantiene, contrariamente a Grimshaw, que la diferencia denotativa no está relacionada con la capacidad argumental de las nominalizaciones sino con la formación derivativa de estas, que se produce en distintos niveles del lenguaje: las nominalizaciones eventivas se generan en la sintaxis y, por tanto, son casos de nominalización sintáctica, mientras que las nominalizaciones resultativas constituyen

Picallo, 1999

casos de nominalización léxica puesto que se derivan al nivel léxico. Respecto a la capacidad argumental, Picallo también considera que los complementos nominales de las nominalizaciones resultativas son argumentales ya que, según ella, se comportan como argumentos reales en lo que concierne a muchos fenómenos gramaticales: pueden ser antecedentes de expresiones anafóricas, pueden ser sujetos de expresiones predicativas y se pueden establecer relaciones interpretativas típicamente argumentales entre los complementos y el núcleo nominal.

A pesar de la discrepancia de estas dos autoras con Grimshaw respecto a la capacidad argumental de las nominalizaciones resultativas, sí que están de acuerdo con ella en la consideración de las nominalizaciones de evento (o proceso) y resultado como unidades léxicas diferentes. Cabe recordar que la representación léxica de ambas denotaciones como unidades léxicas diferentes o sentidos de una misma unidad léxica es un aspecto también controvertido entre los diferentes autores. Desde la corriente lingüística de la GG, de la que todas las autoras arriba reseñadas forman parte, se considera que estas dos denotaciones constituyen unidades léxicas diferentes. Desde otros enfoques teóricos se ha considerado que las dos denotaciones son sentidos de una misma unidad léxica, es decir, se tratan como casos de polisemia. Por ejemplo, Alonso (2004), que aplica la Teoría Sentido-Texto (*Meaning-Text Theory*, MTT en adelante) al estudio del español, afirma que estas nominalizaciones son unidades en las que existe una disyunción de significado; desde la teoría del Lexicón Generativo (*Generative Lexicon*, GL en adelante), Pustejovsky (1995) mantiene que las nominalizaciones son unidades léxicas infraespecificadas; mientras que en otros trabajos (Badia, 2002) se afirma que son, simplemente, unidades léxicas con sentidos distintos.

Meaning-Text Theory
Alonso, 2004

En Alonso (2004) se argumenta que hay sustantivos que presentan una disyunción en su significado puesto que algunos lemas nominales pueden actualizar la lectura eventiva y resultativa en la misma frase sin afectar ello a la comprensión de los enunciados. Por ejemplo, en el caso de (3) el nombre ‘declaración’ se interpreta como un evento y un resultado al mismo tiempo. Únicamente se puede especificar el momento de inicio de un evento, y únicamente de un resultado se puede decir que ocupa cinco páginas. Estos dos significados se originan en la misma unidad léxica, que incluye ambos sentidos (evento y resultado), y es el contexto el que los actualiza a los dos.

- (3) La **declaración** que el juez tomó al testigo, y que comenzó a las once, ocupa cinco folios ⁶.

Respecto a la capacidad argumental, para Alonso todos los nombres que participan en construcciones de verbo soporte seleccionan argumentos. Desde su punto de vista, se puede inferir que si un nombre resultativo participa en una construcción

⁶Este ejemplo se ha extraído de Alonso (2004).

2. NOMINALIZACIONES DEVERBALES: ESTADO DE LA CUESTIÓN

de verbo soporte, tendrá estructura argumental. En español es posible encontrar nombres resultativos en este tipo de estructuras, por ejemplo, ‘hacer acusaciones’, por lo que esta autora argumenta que tanto las nominalizaciones de resultado como las eventivas tienen también estructura argumental.

Generative Lexicon

Pustejovsky, 1995

En el modelo del GL, Pustejovsky (1995) da cuenta de la ambigüedad de las nominalizaciones de proceso (equivalente a nominalizaciones de evento en otros autores) y resultado mediante una representación léxica infraespecificada que denomina *dot-object*. Argumenta que los nombres de proceso-resultado son casos de polisemia complementaria: “*both senses of a logically polysemous noun seem relevant for the interpretation of the noun in the context, but one sense seems ‘focused’ for purposes of a particular context*” (Pustejovsky, 1995, p.31). Así, mantiene que el nombre de proceso-resultado es una unidad léxica compleja que abarca los dos sentidos, que pueden manifestarse conjuntamente o por separado en función del contexto.

En el GL cada sentido de cada palabra se estructura en cuatro ejes de representación: estructura argumental, estructura eventiva, estructura de qualia y estructura de herencia. En Pustejovsky (1995) se describe cómo puede variar la interpretación de los sustantivos (en general) de acuerdo con las primeras tres dimensiones expuestas⁷. De la primera depende el número de argumentos que los sustantivos pueden seleccionar; de la segunda, a qué tipo de eventos se refiere el nombre explícita o implícitamente; y de la tercera, cuál es la fuerza predicativa básica del nominal. En el caso específico de las nominalizaciones de proceso-resultado, la estructura eventiva adquiere una relevancia especial para su interpretación: una nominalización deverbal de proceso-resultado es una unidad compleja, una representación *dot-object* que tiene dos subeventos (un proceso y un resultado) en su estructura eventiva, y esos dos subeventos están relacionados por una relación de precedencia (*Restr*), que determina que el proceso siempre precede al resultado. Según cuál de los dos subeventos se actualice en un determinado contexto esa será la denotación, aunque también puede ocurrir que la denotación quede infraespecificada.

Sobre el concepto de resultado, Pustejovsky plantea que para las nominalizaciones derivadas de verbos de creación (‘construcción’ o ‘desarrollo’) la interpretación de resultado puede corresponder tanto al objeto creado como resultado de la acción, como al estado resultante (Pustejovsky, 1995, p. 172). Sin embargo, también en el marco teórico del GL, Jezek and Melloni (2009) postulan para el italiano que la noción de resultado en las nominalizaciones de verbos de creación (‘construcción’) y de redescrición (‘traducción’) solo puede ser la de objeto-resultado

Jezek and Melloni, 2009

⁷La estructura de herencia identifica cómo una estructura léxica se relaciona con otras estructuras léxicas, es decir, su contribución se centra más en la organización global del léxico que no en los ítems léxicos en sí.

(y no la de estado-resultado). Esta teoría no es la única dónde se distingue entre posibles tipos de resultado. Como veremos, existen propuestas de clasificación de las nominalizaciones en las que la distinción entre evento y resultado es más fina, en las que se subdivide en distinciones más específicas y que afectan especialmente a las nominalizaciones de la clase resultado.

Barque et al., 2009

También en el marco teórico del GL, Barque et al. (2009) identifican para el francés cinco tipos de nominalizaciones deverbales en función básicamente del tipo de verbo del que derivan y de cuatro criterios aspectuales que normalmente se usan para diferenciar las clases aspectuales verbales: dinamicidad, limitación, culminación y duratividad. Los cinco tipos de nominalizaciones deverbales del francés se derivan de los cuatro tipos aspectuales básicos propuestos para los verbos en Vendler (1967): las nominalizaciones de hábito y de proceso se derivan de verbos de actividades, y las nominalizaciones de estado, logro y realización se corresponden con las clases verbales aspectuales de Vendler del mismo nombre. Las nominalizaciones de estado derivan de verbos estativos y al igual que ellos se caracterizan por no ser dinámicas y componerse de un único subevento *State*, que es el núcleo de la estructura argumental del sustantivo (*croyance*, ‘creencia’). Las nominalizaciones de hábito son dinámicas (derivan de verbos de actividades) pero no tienen una limitación en el tiempo, por lo que el único subevento, que es el núcleo de la estructura argumental de la nominalizaciones, es *Process* y se caracteriza por tener una interpretación habitual y un estatus incontable. Según estos autores, esta nueva categoría aspectual en el dominio nominal respecto al verbal, los hábitos, puede corresponderse con una lexicalización de un significado gramatical, que puede ser expresado en el dominio verbal con el verbo ‘soler’ (*jardinage*, ‘jardinería’, se correspondería con el significado verbal de ‘soler dedicarse a las plantas’). En cuanto a las nominalizaciones de proceso, también tienen un único subevento *Process*, que es el núcleo de la estructura argumental de la nominalización, pero las de proceso son dinámicas y limitadas en el tiempo, por lo que son individualizadas y contables (*promenade*, ‘paseo’ puede contabilizarse ‘un paseo’). Las nominalizaciones de logro se caracterizan por ser transiciones dinámicas, limitadas en el tiempo, culminativas y durativas, que se componen de dos subeventos *Process* y *State* en las que el *State* es el núcleo de la estructura argumental de la nominalización (*découvert*, ‘descubrimiento’). Las nominalizaciones realizaciones se caracterizan por ser transiciones dinámicas, limitadas en el tiempo, culminativas y no durativas que se componen de dos subeventos *Process* y *State* en las que el *Process* es el núcleo de la estructura argumental de la nominalización (*réparation*, ‘reparación’). La diferencia entre las nominalizaciones que denotan logros y las que denotan realizaciones es que las primeras ponen el acento en el estado final y, por lo tanto, son durativas, mientras que las segundas focalizan en el proceso y por eso no indican una duración determinada.

Fuera del marco del GL aunque de manera similar a Pustejovsky (1995), Badia (2002), en su trabajo sobre los complementos nominales del catalán, establece que las nominalizaciones de evento y resultado tienen ambas capacidad argumental, pero las segundas no son sustantivos predicativos (equivalentes semánticamente a verbos) mientras que las primeras sí. Además, asegura que la interpretación de evento y resultado corresponde a sentidos diferentes de una misma nominalización, coincidiendo en este aspecto con Pustejovsky (1995) y Alonso (2004). En lo que se refiere a las nominalizaciones resultativas, Badia argumenta que pueden tener dos significados diferentes en función de su capacidad para expresar el complemento objeto del verbo base. Así por ejemplo, en una oración como (4), la nominalización resultativa ‘traducción’ se interpreta como el objeto resultante de la acción del verbo base ‘traducir’, mientras que en (5), la nominalización resultativa ‘análisis’ denota el resultado de la acción del verbo y no el objeto resultante de aquella acción. Badia 2002

(4) [La **traducción**]_{SN} es muy buena.

(5) [El **análisis** [de sangre]_{SP}]_{SN} no mostró ningún peligro.

Una diferencia similar mantiene Levi (1978) quien distingue cuatro tipos de nominalizaciones para el inglés: de acción, de agente, de producto y de paciente. Levi 1978 Las nominalizaciones de acción (*parental refusal*, ‘rechazo paterno’) corresponden a la noción de evento, es decir, denotan una acción (en el ejemplo, la acción de rechazo por parte de los padres). Las de agente (*financial analyst*, ‘analista financiero’) denotan el agente de la acción (la persona que analiza finanzas), y en inglés, como también en español, emplean una gama diferente de sufijos en su proceso derivativo. Las nominalizaciones de producto (*human error*, ‘error humano’) denotan el resultado de una acción (en el ejemplo, lo que es producido por el acto humano de errar) mientras que las de paciente (*students inventions*, ‘inventos de estudiantes’) denotan el objeto resultante de una acción (en el ejemplo, la cosa que los estudiantes han inventado). La diferencia que Levi establece entre las nominalizaciones de producto y de paciente es muy similar a la distinción entre nombres resultativos que Badia presenta: ‘análisis’ en el ejemplo (5) se correspondería con una nominalización de producto mientras que ‘traducción’ en el ejemplo (4) con una nominalización de paciente. En inglés resulta más fácil hacer esta diferenciación puesto que existen sufijos que se especializan en alguna de las dos lecturas (por ejemplo el sufijo *-ee* en *employee*, ‘empleado’ se especializa en la lectura paciente), pero en español la morfología tiende a unir la forma de las nominalizaciones de acción, de producto y de paciente y, por lo tanto, la diferencia es más difícil de establecer, sobre todo entre la de producto y paciente, en las que la distinción es muy sutil: resultado de la acción y objeto resultante.

Eberle et al., 2011 Además de las tipologías que postulan un desdoblamiento en la clase de resultado, existen otras tipologías denotativas en las que se proponen más de dos clases. En Eberle et al. (2011) se analizan las nominalizaciones deverbales formadas con el sufijo *-ung* del alemán, que por su productividad, contenido semántico y denotación podría ser equivalente al sufijo *-ción* del español, y mantienen que estas nominalizaciones pueden llegar a denotar un evento (*messung*, ‘medición’), un estado (*teilung*, ‘división’) y un objeto-resultado (*lieferung*, ‘suministro’). Aclaran que no todas las nominalizaciones en *-ung* son ambiguas por partida triple, sino que dependiendo de la clase semántica del verbo base la nominalización podrá tener tres, dos o solo una de las tres denotaciones posibles.

Balvet et al., 2010 En Balvet et al. (2010), un trabajo sobre las nominalizaciones deverbales del francés, se distinguen cuatro tipos de nominalizaciones: estados (*admiration*, ‘admiración’), eventos durativos (*opération*, ‘operación’), eventos puntuales (*explosion*, ‘explosión’) y objetos (*bâtiment*, ‘edificio’). Estos mismos autores (Balvet et al., 2011) han refinado aún más su tipología y distinguen hasta 11 tipos de nominalizaciones deverbales. Estas 11 clases se componen de cuatro clases nominales paralelas a las clases aspectuales de los verbos de Vendler: estados (*admiration*, ‘admiración’), actividades (*promenade*, ‘paseo’), realizaciones (*déménagement*, ‘mudanza’), logros (*acquisition*, ‘adquisición’); de dos clases específicas del dominio nominal: los sustantivos objetos (*construction*, ‘construcción/edificio’), es decir, sustantivos que designan el objeto resultante de la acción verbal, y sustantivos hábitos (*jardinage*, ‘jardinería’), que son sustantivos que expresan un hábito; y de cinco clases complejas que resultan de combinar dos de las seis clases anteriores: los sustantivos logros-estados (*emprisonnement*, ‘encarcelamiento’), realizaciones-estados (*invasion*, ‘invasión’), actividades-estados (*rétrécissement*, ‘estrechamiento’), realizaciones-objetos y logros-objetos, para las que no se proporcionan ejemplos.

Role Reference Grammar Desde la Gramática del Rol y la Referencia (*Role Reference Grammar*, RRG en adelante) también se ha trabajado sobre las nominalizaciones y su denotación. Para el inglés, Nunes (1993) establece cinco tipos de nominalizaciones deverbales: de proceso, que denotan la acción significada por el verbo base (6); de resultado, que denotan una nueva creación resultante del verbo base (7); de acción acumulada, que denotan la suma total de actividad de un verbo (8); de estados experimentales, nominalizaciones de verbos estativos o nominalizaciones derivadas de un estado presente en un verbo determinado (9); y estados-experimentales resultativos, es decir la contrapartida resultativa a la clase previa (10).

- Nunes, 1993**
- (6) [The [documents’]_{SN} **destruction** [by the North]_{SP}]_{SN}.
[La **destrucción** [de los documentos]_{SP} [por parte del Norte] _{SP}]_{SN}.
- (7) [The **invention**]_{SN} was put on display.
[El **invento**]_{SN} fue puesto en marcha.

2. NOMINALIZACIONES DEVERBALES: ESTADO DE LA CUESTIÓN

- (8) [The **attack**]_{SN} was unexpected.
*[El **ataque**]_{SN} fue inesperado.*
- (9) [[Sam's]_{SN} **interest** [in math]_{SP}]_{SN}.
*[El **interés** [de Sam]_{SP} [en las matemáticas]_{SP}]_{SN}.*
- (10) Sam has many [**interests**]_{SN}.
*Sam tiene muchos [**intereses**]_{SN}.*

A continuación resumimos en la Tabla 2.1 las distintas tipologías de nominalizaciones deverbales en función de su denotación según los distintos autores reseñados.

Evento y Resultado: Clasificaciones					
Autores	2 Clases	3 Clases	4 Clases	5 Clases	11 Clases
Zubizarreta'87			+		
Grimshaw'90	+				
Borer'97	+				
Gràcia i Solé'95	+				
Martí i Girbau'89	+				
Demonte'89	+				
Picallo'99	+				
Alexiadou'01	+				
Alonso'04	+				
Pustejovsky'95		+			
Barque et al'09				+	
Badia'02		+			
Levi'78		+			
Eberle et al'11		+			
Balvet et al'10			+		
Balvet et al'11					+
Nunes'93				+	

Tabla 2.1: Tabla resumen de las clasificaciones según la denotación de las nominalizaciones deverbales

A pesar de las distintas propuestas de clases denotativas, la clasificación en evento y resultado de las nominalizaciones deverbales es la más extendida entre los distintos autores, por eso es la que nosotros decidimos aplicar en nuestro estudio. A continuación resumimos los criterios (un total de doce) que desde la bibliografía se han propuesto, mayoritariamente aplicados al inglés, para la distinción de las dos denotaciones básicas: evento y resultado. Como avanzamos al principio de este capítulo, aquellos autores que consideran que las dos denotaciones son unidades léxicas distintas son los que más criterios proponen, si bien alguno de los autores que mantienen que son sentidos de una misma unidad léxica también secundan alguno de los criterios. En la Tabla 2.2 presentamos los doce criterios más relevantes usados en la bibliografía y los autores que los proponen.

2.1.1.1. Criterios lingüísticos

1) Clase Verbal. Uno de los criterios más utilizados para determinar la denotación de la nominalización deverbal es la clase de verbo de la que deriva. La mayoría de clasificaciones que hemos visto tienen en cuenta el tipo de verbo del que deriva la nominalización: o bien se tienen en cuenta aspectos más sintácticos como la transitividad-intransitividad, o bien se atiende a aspectos más semánticos como el aspecto del verbo o clases semánticas específicas. Desde un punto de vista más sintáctico, Picallo (1999) para el español y Alexiadou (2001) para el inglés, alemán y griego, mantienen que los verbos inergativos dan lugar siempre a nominalizaciones resultativas, mientras que los inacusativos resultan a menudo en nominalizaciones ambiguas entre ambas lecturas. Respecto a los predicados estativos, se suele afirmar que dan lugar a nominalizaciones con un comportamiento sintáctico similar a las nominalizaciones resultativas (Picallo, 1999). También Zubizarreta (1987) postula que los predicados estativos dan lugar a un tipo concreto de nominalización. En cuanto a los verbos transitivos, Alexiadou (2001) sostiene que pueden dar lugar a nominalizaciones únicamente eventivas o a nominalizaciones ambiguas entre la lectura resultativa y la eventiva. Picallo (1999), por su parte, mantiene que los verbos transitivos pueden derivar tanto nominalizaciones eventivas como resultativas inequívocamente si se dan las condiciones sintácticas requeridas; si no, las nominalizaciones derivadas de verbos transitivos tienen una interpretación ambigua entre ambas lecturas. Desde un punto de vista más semántico, en los trabajos de Balvet et al. (2010, 2011) se tiene en cuenta la clase aspectual de los verbos base para establecer la clase denotativa de las nominalizaciones. También en el marco del GL se tiene en cuenta la clase semántica del verbo base para distinguir entre los dos tipos de resultados (resultado de la acción, objeto resultante) ya que se mantiene que los verbos de creación y de redescrición solo pueden dar lugar a objetos resultantes.

Criterios	Zubizarreta	Grimshaw	Alexiadou	Picallo	Alonso	Badia	Balvet et al.	GL
Clase Verbal	+	-	+	+	-	+	+	+
Pluralización	+	+	-	-	+	-	-	-
Determinante	-	+	-	+	+	-	-	-
Prep + Agente	-	-	-	+	-	+	-	-
Obligatoriedad Arg. Int.	-	+	-	+	-	-	-	-
Poseedores vs. Arg.	-	+	+	-	-	-	-	-
Predicado Verbal	-	+	-	+	-	+	-	-
Mod. Aspectuales	-	+	+	+	-	-	-	-
Estructuras Control	-	+	-	+	-	-	-	-
Mod. del agente	-	+	-	-	-	-	-	-
Afectación objeto	-	-	+	-	-	-	-	-
Telicidad/Atelicidad	-	-	+	-	-	-	-	-

Tabla 2.2: Criterios Lingüísticos para la distinción Evento vs. Resultado. Leyenda: Arg. Int. en la quinta fila significa Argumento Interno, Arg. en la sexta fila, argumentos y Mod. en la fila ocho y diez, modificadores.

2) Capacidad de pluralización. Uno de los rasgos que según los autores identifica más claramente a las nominalizaciones resultativas del inglés (Zubizarreta, 1987; Grimshaw, 1990) y del español (Picallo, 1999; Alonso, 2004) es su capacidad de pluralización. La mayoría considera que las resultativas pueden aparecer en plural, a diferencia de las nominalizaciones eventivas que suelen aparecer siempre en singular.

3) Tipo de determinante. En la bibliografía sobre las nominalizaciones del inglés (Grimshaw, 1990; Alexiadou, 2001) y del español (Picallo, 1999; Alonso, 2004) es comúnmente aceptado que las nominalizaciones eventivas solo aparecen con el artículo definido, mientras que las resultativas se caracterizan por admitir todo tipo de determinantes: definido, indefinido, demostrativos, numerales, etc.

4) Preposición + Agente. En las nominalizaciones del español que derivan de verbos transitivos, se considera que la preposición que introduce el complemento agentivo puede determinar la denotación de la nominalización. Picallo (1999) afirma que un complemento agentivo introducido por la preposición *de* implica una lectura resultativa del nominal, mientras que si la preposición es *por*, o la locución prepositiva *por parte de*, la nominalización tiene una lectura eventiva. Badia (2002) sostiene lo mismo para el catalán.

5) Obligatoriedad del argumento interno. Este criterio, expuesto por Picallo (1999) para el español y por Grimshaw (1990) para el inglés, establece que solo las nominalizaciones eventivas exigen la presencia del argumento interno mientras que en las nominalizaciones resultativas este no es necesario. Badia (2002) argumenta que en catalán no siempre es necesaria la realización de este argumento para obtener una interpretación eventiva de la nominalización. Así por ejemplo, el sustantivo ‘destrucción’ denotaría tanto un evento en la oración ‘La destrucción de la casa por parte de Juan’ como en ‘La destrucción tuvo lugar ayer’.

6) Poseedores vs. argumentos. Uno de los criterios que Grimshaw (1990) postula en inglés para diferenciar entre nominalizaciones resultativas y eventivas es que los sintagmas preposicionales introducidos por la preposición *by* (*by*-SPs), los adjetivos relacionales y los determinantes posesivos se interpretarían como argumentos externos (sujetos) en el caso de las nominalizaciones eventivas. En cambio, estos mismos constituyentes se interpretarían como poseedores, es decir, como no argumentales en el caso de las nominalizaciones resultativas. Otros autores, como Picallo (1999) para el español y Badia (2002) para el catalán, en cambio, mantienen que en estas lenguas los determinantes posesivos pueden ser interpretados como argumentos en ambos tipos de nominalizaciones. Picallo (1999) afirma, a diferencia de Grimshaw, que los adjetivos relacionales solo aparecen como argumentos en las nominalizaciones resultativas. Se trata, por lo tanto,

de un criterio que no se mantiene para las diferentes lenguas según los distintos autores.

7) Predicado Verbal. El tipo de predicado verbal con el que la nominalización se combina puede ser un indicador para determinar el tipo de denotación en inglés (Grimshaw, 1990), en español (Picallo, 1999) y en catalán (Badia, 2002). Las nominalizaciones resultativas parecen combinarse únicamente con predicados atributivos, mientras que las nominalizaciones eventivas serían sujetos de predicados del tipo ‘tener lugar’ u ‘ocurrir’.

8) Modificadores Aspectuales. Autores como Grimshaw (1990), Picallo (1999) y Alexiadou (2001) consideran que los modificadores aspectuales que complementan los predicados verbales son los mismos que aparecerían en las nominalizaciones eventivas, pero nunca ocurrirían en nominalizaciones resultativas. En español, además, en este criterio Picallo (1999) también tiene en cuenta el tipo de preposición que introduce el SP modificador aspectual-temporal: si un SP modificador temporal de un sustantivo deverbal es introducido por la preposición ‘de’, la lectura de la nominalización sería resultativa, mientras que si el modificador temporal es un SN, entonces la interpretación sería eventiva.

9) Estructuras de control. Según Grimshaw (1990) para el inglés y Picallo (1999) para el español solo los nominales eventivos admitirían estructuras de control en oraciones finales de infinitivo. Por ejemplo, en el SN ‘La asignación de problemas fáciles para aprobar a todos los estudiantes’, la oración subordinada final introducida por la preposición ‘para’ es la que da la clave para considerar a la nominalización deverbal ‘asignación’ como eventiva.

10) Modificadores del agente. Grimshaw (1990) mantiene para el inglés que un modificador (por ejemplo, un adjetivo del tipo *intentional*, ‘intencional’ o *voluntary*, ‘voluntario’) que se refiere a un complemento interpretado como agente es un indicador de que dicha nominalización recibe una interpretación eventiva.

11) Afectación del objeto. Alexiadou (2001) sostiene para el inglés y el griego que los predicados verbales transitivos con un objeto afectado (‘destruir’) solo dan lugar a nominales eventivos.

12) Telicidad/Atelicidad. Es también Alexiadou (2001) quien afirma que de los verbos transitivos atélicos solo derivan nominalizaciones resultativas mientras que de los verbos transitivos télicos solo derivan nominalizaciones eventivas.

Estos doce criterios han sido propuestos por los diferentes autores para establecer la diferencia denotativa entre los dos tipos básicos de denotaciones, evento (o proceso) y resultado. A pesar de que, en general, existe un acuerdo amplio entre

los autores respecto a los criterios, no siempre es así, como es el caso de las disparidades respecto a la obligatoriedad del argumento interno o a la consideración de los complementos nominales como argumentos o poseedores. En este sentido, nos parecía necesario la realización de un estudio empírico que contrastase estos criterios con el objetivo de validar cuáles son válidos para el español (Capítulo 5).

2.1.2. Nominalizaciones deverbales y estructura argumental

En esta sección nos centramos en aquellos autores que desde marcos teóricos distintos han estudiado las nominalizaciones deverbales poniendo el acento en la estructura argumental. Por estructura argumental se entiende la representación léxica de la información sintáctico-semántica de un predicado. En concreto, en la estructura argumental se especifica el número de argumentos semánticos requeridos (es decir, los participantes implicados) por la unidad léxica predicativa (en nuestro caso, la nominalización deverbal) y el tipo de relación semántica que dichos argumentos mantienen con el predicado, que normalmente se establece mediante papeles temáticos (agente, paciente, tema, etc.).

Entre los marcos teóricos estudiados se encuentran diferentes propuestas pertenecientes a la familia de las denominadas gramáticas generativas, es decir, teorías lingüísticas que pretenden dar cuenta de la capacidad generativa del lenguaje, de la manera en que cada lengua es capaz de producir el conjunto de oraciones bien formadas. Entre ellas destaca el marco teórico conocido como Gramática Generativa (GG), dominante en el panorama lingüístico desde los años 50 (Chomsky, 1965) hasta los 90 (Chomsky, 1995). Esta se caracteriza por ser una gramática generativa transformacional, esto es, una gramática en la que se postula dos niveles de representación sintáctica, la Estructura Profunda (EP) y la Estructura Superficial (ES), y en la que las transformaciones dan cuenta de una gama amplia de fenómenos, como son la relación entre estructuras activas y pasivas, el tratamiento de elementos interrogativos y, en general, los fenómenos que implican el desplazamiento u omisión de algún constituyente. Los otros dos modelos que revisamos son las denominadas gramáticas generativas de restricciones léxicas: la Gramática Léxico-Funcional (*Lexical Functional Grammar*, LFG en adelante) y la Gramática de Estructura Sintagmática regida por el Núcleo (*Head-driven Phrase Structure Grammar*, HPSG) se caracterizan, en cambio, por ser monoestratales (postulan un único nivel de representación sintáctica) y lexicalistas. Nos centramos, por tanto, en cómo estos marcos teóricos dan cuenta de la estructura argumental en sus teorías y nos fijamos especialmente en aquellos trabajos que tratan las nominalizaciones deverbales. Además de las gramáticas generativas, revisamos también el tratamiento de la estructura argumental desde marcos teóricos como el Lexicón Generativo (*Generative Lexicon*, GL) o la Teoría Sentido-Texto (*Meaning-Text Theory*, MTT).

2. NOMINALIZACIONES DEVERBALES: ESTADO DE LA CUESTIÓN

En la GG, específicamente en el modelo teórico propuesto en la Teoría de la Rección y el Ligamento (Chomsky, 1981) (*Government and Binding Theory*, en adelante GB), la gramática es modular, concretamente se organiza en cuatro módulos autónomos pero relacionados entre sí: el componente léxico; el sintáctico, donde se distingue el nivel de EP y ES relacionados por las transformaciones; el fonológico, encargado de dotar de representación fonética a las oraciones; y el semántico, que se relaciona con el significado y que conecta la facultad del lenguaje con las facultades perceptuales y motoras. En esta teoría, es en el componente léxico donde se especifica la estructura argumental de las unidades léxicas, es decir, que la estructura argumental forma parte de las entradas léxicas en las que también se recoge información acerca de la categoría, la subcategorización y las restricciones selectivas de la unidad léxica. Como se ha dicho anteriormente, en la estructura argumental se especifica el número de argumentos semánticos requeridos por la unidad léxica predicativa y el tipo de relación semántica que dichos argumentos mantienen con el predicado. En Grimshaw (1990) se mantiene que la EP, la estructura sintáctica primaria, se proyecta desde la estructura argumental, de ahí, la importancia de esta representación en la GG.

En los años 70, Chomsky (1970) presentó el artículo germinal sobre las nominalizaciones deverbales en dicho marco teórico. En este trabajo se distingue entre tres tipos de nominalizaciones en inglés: de gerundio (*John's criticizing the book*, 'La crítica del libro por parte de John'), mixtas (*The barbarian's destruction of the city*, 'La destrucción de la ciudad por los bárbaros') y derivadas (*Belushi's mixing of drugs led to his demise*, 'La mezcla de drogas de Belushi le llevó a su fallecimiento'). En el paradigma de la GG de estos años, Chomsky trata de argumentar en este artículo que mientras el primer tipo de nominalización se genera de manera transformacional, a través de operaciones sintácticas, los otros dos tipos lo hacen mediante la extensión de reglas léxicas, es decir, a nivel del léxico y no de la sintaxis. A pesar de que en este trabajo no se trata la denotación ni la estructura argumental de las nominalizaciones deverbales, nos parece necesario referirnos a él puesto que fue el iniciador dentro de la GG de una larga tradición de trabajos sobre las nominalizaciones deverbales (Zubizarreta, 1987; Grimshaw, 1990; Picallo, 1999; Alexiadou, 2001). Como vimos en la sección anterior (Sección 2.1.1), existe un grupo de autores generativistas (Grimshaw, 1990; Borer, 1997; Gràcia i Solé, 1995; Demonte, 1989; Martí i Girbau, 2002) que afirman que la diferencia entre nominalización de evento (o proceso) y resultado viene dada por la presencia de estructura argumental en las primeras y la carencia de estructura argumental en las segundas, mientras que otros autores de esta misma corriente (Picallo, 1999; Alexiadou, 2001) afirman que tanto nominalizaciones de evento como de resultado tienen estructura argumental dado que consideran que la diferencia entre ambas denotaciones estriba en el diferente proceso de derivación y en la diferente proyección funcional, respectivamente.

Gramática Generativa
Transformacional

Government and Binding,
GB

Chomsky, 1970

Gramáticas Generativas de Restricciones Léxicas La LFG y la HPSG son gramáticas generativas no transformacionales, mono-estratales, esencialmente lexicalistas, de ahí llamadas gramáticas de restricciones léxicas, en las que se otorga una importancia extraordinaria al componente léxico, módulo a partir del cual se proyecta la información sintáctica y semántica. Recuérdese que este tipo de gramáticas se diferencian de la GG en que solo existe un único nivel de análisis sintáctico y el concepto de transformación ya no es necesario.

Lexical-Functional Grammar, LFG La gramática LFG (Bresnan, 1982) también se organiza de manera modular y parte del léxico como componente básico a partir del cual toman la información los dos niveles de descripción sintáctica que esta teoría asigna a toda oración de la lengua: la estructura de constituyentes y la estructura funcional. En la primera se especifican las configuraciones sintagmáticas (relaciones de dominio y precedencia de las palabras y los sintagmas) y en la segunda se representan las funciones gramaticales (sujeto, objeto directo, objeto indirecto, etc.) y se especifica la información interpretable semánticamente. El componente léxico por su parte incluye el conjunto de entradas léxicas y una serie de reglas léxicas que sirven para establecer las relaciones sistemáticas entre dos estructuras sintáctico-semánticas (como la activa-pasiva, por ejemplo). En las entradas léxicas se especifica la estructura argumental de los predicados, además de la forma léxica, la categoría y los rasgos morfosintácticos. Toda esta información se representa en forma de una estructura de rasgos, que representan parejas de atributo-valor. En esta teoría, la relación entre la estructura argumental y la estructura sintáctica se establece a través de las funciones gramaticales (*grammatical functions*), que constituyen categorías primitivas de la gramática, lo que supone también una diferencia respecto a la GGT. Se distinguen dos tipos de funciones gramaticales: funciones gramaticales no restringidas semánticamente (SUJ, OBJ) y funciones gramaticales restringidas semánticamente (OBL OBJ). La relación entre argumentos y funciones gramaticales se define mediante reglas de enlace que especifican qué funciones gramaticales pueden realizar los distintos argumentos.

Rappaport, 1983 Rappaport (1983) afirma que las nominalizaciones deverbales, pese a heredar la estructura argumental del verbo del cual derivan, no utilizan las funciones de SUJ y OBJ. Argumenta que únicamente tienen a su disposición la función gramatical POSS (possessive) que es propia de los sintagmas nominales y la serie de funciones oblicuas (OBL OBJ) cuyos argumentos están restringidos semánticamente por la preposición que las introduce. En definitiva, las reglas de enlace entre argumentos y funciones sintácticas son diferentes en el dominio verbal y en el nominal, pero se reconoce que las nominalizaciones deverbales al igual que los verbos de las que derivan, poseen estructura argumental.

Meinschaefer, 2005 También en el marco de la LFG, Meinschaefer (2005) se centra en las nominalizaciones deverbales del español. Propone tres funciones gramaticales para las nominalizaciones: la función POSS, la función TOPPOSS y la función OBL.

La función POSS en español se realiza mediante un sintagma preposicional (SP) introducido por la preposición *de* y especifica un argumento tema, un argumento meta o un argumento agente. La función TOPOSS, que se corresponde con el determinante posesivo español, también codifica estos tres argumentos, pero estos deben además estar marcados como “información ya proporcionada”, es decir, información ya conocida. La función OBL solo puede realizar argumentos de tipo meta o agente y en español se corresponde con SPs introducidos por preposiciones distintas a *de*.

La HPSG (Pollard and Sag, 1987, 1994) concede gran importancia a los núcleos léxicos ya que a partir de la información que contienen se proyecta gran parte de la información a nivel sintáctico. Este modelo gramatical se organiza básicamente en un componente léxico, que incluye las entradas léxicas o signos léxicos y las reglas léxicas (de derivación, composición, alternancia de diátesis, etc.), un conjunto finito de reglas gramaticales y una serie de principios (de subcategorización, de rasgos de núcleo, etc.).

Badia and Saurí (2008) desde la HPSG también secundan la presencia de estructura argumental en las nominalizaciones deverbales. En el signo léxico, los argumentos de las nominalizaciones se representarían en el nivel semántico. Estos autores distinguen tres tipos de argumentos inspirándose en el GL de Pustejovsky: argumentos verdaderos (*true arguments*), argumentos por omisión (*default arguments*) y argumentos a la sombra (*shadow arguments*) (Pustejovsky, 1995). Por argumento verdadero se entiende un participante subcategorizado, requerido sintácticamente por el predicado (‘la casa’ en ‘Juan construyó la casa’). Los argumentos por omisión son casos de argumentos que son necesarios semánticamente para la interpretación del predicado pero que no se requieren sintácticamente (‘a la estación’ en ‘Juan entró a la estación’). Los argumentos en la sombra se caracterizan por estar incorporados al ítem léxico (‘una canción’ en ‘Juan cantó’). Finalmente, los adjuntos verdaderos no están ligados a ningún ítem léxico particular sino que forman parte de la interpretación de la situación del mismo, es decir, se corresponde básicamente con las expresiones de espacio y tiempo en las que casi cualquier predicado se ubica (‘el martes’ en ‘Juan llegó tarde el martes’). Sin embargo, según Badia y Saurí, los complementos de las nominalizaciones son en su mayoría opcionales aunque seleccionados semánticamente por el núcleo léxico. Para explicar dicha opcionalidad, etiquetan los complementos opcionales con un rasgo específico y asumen que la lista de argumentos contiene información sobre su semántica. Esta lista de argumetos (*argstr*) permite mantener la información semántica de estos complementos incluso si esta está ausente de la cadena superficial. Por ejemplo, en la nominalización deverbal ‘construcción’, los argumentos agente, material y resultado son opcionales (argumentos por defecto) y se representan como tales en las listas de valencias (nivel sintáctico). Las nominalizaciones resultativas, marcadas aún más si cabe por la opcionalidad de los argumentos,

también tendrían argumentos por defecto en su estructura argumental, como pasa con las nominalizaciones eventivas.

Generative Lexicon, GL
Pustejovsky, 1995

En el modelo del GL (Pustejovsky, 1995) cada sentido de cada palabra se estructura en cuatro niveles de representación: estructura argumental, estructura eventiva, estructura de qualia y estructura de herencia. Si bien la denotación se representa en la estructura eventiva (véase la Sección 2.1.1), los argumentos se representan en la estructura argumental. En este marco teórico se distinguen cuatro tipos de argumentos: argumentos verdaderos, argumentos por omisión, argumentos en la sombra y adjuntos verdaderos. Las nominalizaciones deverbales, que como vimos constituyen en este marco un *dot-object*, también pueden tener en su estructura argumental estos cuatro tipos de argumentos.

Meaning-Text Theory,
MTT

Mel'cuk, 1984

En la MTT de Igor Mel'cuk (Mel'cuk, 1981) también se mantiene que las nominalizaciones deverbales tienen capacidad argumental. En el *Dictionnaire Explicative et Combinatoire* (DEC, en adelante) (Mel'cuk et al., 1984), se representan las nominalizaciones del francés y para cada uno de sus sentidos se incluye una definición semántica de la nominalización en la que el uso de variables explicita los actantes semánticos, que en la MTT son equivalentes al concepto de argumentos. Por ejemplo, en 'La promesa de X a Y de Z', las variables X, Y, Z representan los argumentos y, por tanto, los actantes semánticos de 'promesa'. Estas mismas variables sirven también para indicar los actantes sintácticos, que no son más que actantes semánticos que ocupan una posición privilegiada en el esquema de régimen, el esquema sintáctico de cada sentido. Cada acepción se completa con información acerca de las posibles combinaciones léxicas de la nominalización y de cómo se realizan sintácticamente. Las combinaciones léxicas se expresan a través de funciones léxicas (existen 50 funciones léxicas diferentes) que proporcionan todas las coocurrencias léxicas idiomáticas del lexema. En las nominalizaciones, las funciones léxicas más relevantes desde nuestra perspectiva son Vj y Oper. La primera (Vj) relaciona las nominalizaciones con los verbos de los cuales derivan. La segunda (Oper) se refiere a los verbos semánticamente vacíos con los que el nombre se combina y que adquieren el actante semántico de este. Esto hace referencia a las construcciones con verbos de soporte como 'tomar una decisión' en la que el verbo está desemantizado y es el sustantivo el que selecciona los actantes semánticos de la construcción.

Como conclusión cabe decir que la mayoría de los marcos teóricos presentados (LFG, HPSG, GL, MTT) consideran que las nominalizaciones, al igual que los verbos, pueden tener estructura argumental independientemente de su denotación. Solamente en el marco de la GG existe disparidad de opiniones entre diferentes autores: mientras que Grimshaw (1990) y Borer (1997) consideran que solo las nominalizaciones eventivas tienen estructura argumental, y no las resultativas, Picallo (1999) y Alexiadou (2001) establecen que ambos tipos de nominalizaciones tienen estructura argumental.

2.2. Aproximaciones Computacionales

En esta sección se revisan los trabajos que desde un punto de vista computacional se han centrado en el tratamiento de las nominalizaciones deverbales. Primero, nos detendremos en los recursos lingüísticos que representan las nominalizaciones deverbales para diferentes lenguas, comparando la información que consideran fundamental en su representación léxica (Sección 2.2.1). A continuación, repasaremos los distintos sistemas automáticos relacionados con el tratamiento de las nominalizaciones deverbales, haciendo especial hincapié en aquellos que tratan la denotación o la estructura argumental con la intención de establecer las comparaciones pertinentes con los sistemas que se han desarrollado en el marco de este trabajo (Sección 2.2.2).

2.2.1. Recursos

Prácticamente todos los sistemas automáticos de tratamiento de las nominalizaciones deverbales se apoyan en la disponibilidad y uso de recursos lingüísticos pertinentes. De ahí que comencemos nuestro análisis describiendo estos recursos.

En el marco de la lingüística computacional existen diferentes recursos lingüísticos que representan las nominalizaciones deverbales: léxicos, bases de datos, ontologías y corpus. Sin embargo, no todos contienen la misma información. Por ejemplo, en el léxico NOMLEX (Macleod et al., 1998) y en el corpus NomBank (Meyers et al., 2004b; Meyers, 2007) la denotación de las nominalizaciones deverbales no se tiene en cuenta, mientras que en WordNet⁸ (Fellbaum, 1998) la estructura argumental de las nominalizaciones no se representa. Los primeros están más interesados en la representación del significado de las proposiciones, es decir, en la representación de la estructura argumental, mientras que WordNet se centra en la representación del significado léxico. En este recurso los sentidos léxicos se definen a través de las relaciones con otros sentidos. La mayoría de los recursos que describimos a continuación son recursos monolingües que se centran principalmente en el inglés, sin embargo hay también recursos dedicados a otras lenguas como el francés (Balvet et al., 2010), el alemán (Burchardt et al., 2009), el japonés (Ohara, 2009), el ruso (Spencer and Zaretskaya, 1999) o el español (Subirats, 2009). Entre todos los recursos que representan las nominalizaciones, el único multilingüe, que nosotros conozcamos, es el corpus paralelo para el danés, inglés, alemán, italiano y español presentado en el trabajo de Hoeg Muller (2010). Además, la mayoría de los recursos que presentamos han sido creados de manera manual, a excepción de NOMLEX-PLUS (Meyers et al., 2004a), lo cual consti-

⁸La consideración de WordNet como una ontología o como una base de datos léxico-conceptual varía entre diferentes autores, no hay unanimidad.

tuye una diferencia fundamental con el léxico AnCora-Nom, desarrollado en el marco de este trabajo, tal y como veremos en el Capítulo 9.

A continuación presentamos los diferentes recursos que recogen la denotación y/o la estructura argumental en su representación de las nominalizaciones deverbales.

NOMLEX

NOMLEX⁹ (Macleod et al., 1998) es un léxico de nominalizaciones deverbales derivadas morfológicamente del inglés que contiene 1.025 entradas léxicas creadas manualmente. Este recurso no solo describe los complementos que una nominalización puede tener (indicando el tipo de constituyente) sino que también asocia a esos complementos los argumentos del verbo base, esto es, indica para cada complemento nominal el complemento verbal con el que se correspondería (sujeto, objeto directo, objeto indirecto, etc.). En NOMLEX se distinguen cuatro tipos de nominalizaciones: 1) nominalizaciones *verb-nom*, que están derivadas de un verbo (*to destroy, destruction*; ‘destruir’, ‘destrucción’); 2) nominalizaciones *verb-part*, que también se derivan de un verbo pero tienen la particularidad que incorporan una partícula prepositiva (*to take over, takeover*; ‘absorber’, ‘absorción’); 3) nominalizaciones *subj-nom*, que denotan el sujeto del verbo base (*to teach, teacher*; ‘enseñar’, ‘enseñante’); y 4) nominalizaciones *obj-nom*, que denotan el objeto del verbo base (*to employ, employee*; ‘emplear’, ‘empleado’). A partir de NOMLEX se creó posteriormente de manera automática el léxico NOMLEX-PLUS (Meyers et al., 2004a) en el que se incrementa el número de entradas léxicas hasta 7.050 tras incluir nominalizaciones deadjetivales, nominalizaciones deverbales *cousin*, es decir, no derivadas morfológicamente (‘éxito’, ‘tener éxito’), y otros tipos de sustantivos con capacidad argumental como los sustantivos relacionales (‘hermano’).

NomBank

En el proyecto NomBank (Meyers et al., 2004b; Meyers, 2007) se llevó a cabo manualmente la anotación semántica de la estructura argumental de todos los sustantivos del corpus PennTreeBank (1 millón de palabras) (Palmer et al., 2005), entre los que se incluían las nominalizaciones deverbales. NomBank comparte el esquema de anotación del proyecto PropBank (Palmer et al., 2005), en el que se realizó la anotación de la estructura argumental de los predicados verbales del mismo corpus. Los argumentos siguen un orden numérico incremental—arg0, arg1, arg2, arg3, arg4— que expresa el grado de proximidad del argumento con respecto a su predicado y los adjuntos se etiquetan como argM. Estas etiquetas son abstractas si las comparamos con los papeles temáticos más específicos usados en VerbNet (Kipper et al., 2000) y (Kipper et al., 2006) o más aún en FrameNet (Baker et al., 1998) y (Ruppenhofer et al., 2006). Sin embargo, cabe mencionar que el grado de abstracción en NomBank y PropBank viene dado porque en estos proyectos se considera que los argumentos se definen en base a su lexema, es decir,

⁹<http://nlp.cs.nyu.edu/nomlex/index.html>

2. NOMINALIZACIONES DEVERBALES: ESTADO DE LA CUESTIÓN

que se especifican para cada unidad predicativa a partir de las etiquetas numéricas más generales.

En la línea de PropBank y NomBank aunque con un esquema más amplío, recientemente se han desarrollado los proyectos TimeBank (Pustejovsky et al., 2005) y FactBank (Saurí and Pustejovsky, 2009). En el primero se anotan las marcas temporales de los textos que conforman TimeBank (183 documentos) y el tipo de relación que estas mantienen con los eventos de aquel texto (7.935 eventos), que también son anotados. En el segundo, FactBank, se marcan las expresiones que especifican el grado de certeza o veracidad de un evento del texto, que también son anotados (en total se anotan 9.488 eventos). Aunque son fenómenos alejados de nuestro objeto de estudio, si nos parece interesante destacar aquí que en ambos proyectos se considera que los sustantivos deverbales denotan eventos y se anotan como tales.

TimeBank y FactBank

Los recursos lingüísticos que se están creando de manera manual para varias lenguas en el marco del proyecto FrameNet¹⁰ es otra de las grandes propuestas de representación léxica que incluye las nominalizaciones deverbales. Este proyecto está basado en la teoría de la Semántica de Marcos, *Frame Semantics* (Fillmore, 1976) que a su vez se basa en la Gramática de Casos, *Case Grammar* (Fillmore, 1968), y se respalda en la evidencia de un corpus real, el *British National Corpus*¹¹ (Aston and Burnard, 1998). Su objetivo es documentar el rango de combinaciones sintácticas y semánticas posibles (valencias) para las palabras predicativas, que incluyen verbos, sustantivos y adjetivos. Construyen marcos semánticos mediante la anotación de un conjunto de ejemplos para cada predicado y mediante la descripción de la red de relaciones entre los diferentes marcos así creados. Cada uno de estos marcos semánticos contiene los elementos correspondientes al marco de la palabra objeto *target* (similares a los papeles temáticos) y sus realizaciones sintácticas correspondientes, incluyendo información sobre las funciones gramaticales y los tipos de sintagma (SN, SP...). Cabe destacar que los elementos de marco son específicos para cada marco por lo que no existe el grado de generalización que se daba en el esquema de anotación de NomBank con los argumentos numerados. Si en aquel esquema los argumentos numerados eran generales para todos los predicados y solo se revelaban específicos en contacto con cada lexema, en FrameNet los elementos del marco son suficientemente específicos y no necesitan ser interpretados junto a su lexema (el agente de ‘construir’, se interpreta como el elemento constructor del marco semántico de ‘construir’). El uso de papeles temáticos es útil para representar el significado proposicional y para dar cuenta de las relaciones de significado sistemáticas entre estructuras sintáctico-semánticas (alternancias de diátesis). Sin embargo, definir un conjunto estándar de papeles

FrameNet

¹⁰<http://framenet.icsi.berkeley.edu/>

¹¹<http://www.natcorp.ox.ac.uk/>

temáticos es problemático. PropBank y FrameNet son dos aproximaciones diferentes a este problema. En PropBank se ha apostado por una representación más general, no ligada a ninguna teoría, que permite el uso de las mismas etiquetas para diferentes predicados. Esta propuesta favorece el rendimiento de sistemas de SRL, por ejemplo, que tienen más datos sobre los que aprender cada etiqueta. Por su parte, en FrameNet los papeles son más específicos y están ligados a la teoría de la Semántica de Marcos, por lo que resultan más informativos desde un punto de vista lingüístico.

En lo que se refiere a las nominalizaciones deverbales, en FrameNet se clasifican en eventivas (*replacement* ‘reemplazamiento’) o en entidades (*building* ‘construcción’), diferenciación similar a la distinción entre evento y resultado. Las nominalizaciones eventivas se representan en el marco semántico del verbo base mientras que las de entidad pertenecen a otro marco semántico. En este sentido, podemos decir que los dos tipos de denotaciones constituyen dos unidades léxicas diferentes. Para el inglés existe un recurso en línea con 11.600 unidades léxicas. Además del inglés, existen propuestas de FrameNets para otras lenguas como el alemán¹² (Burchardt et al., 2009), el japonés¹³ (Ohara, 2009) o el español¹⁴ (Subirats, 2009). El FrameNet español contiene 1.200 unidades léxicas repartidas en poco más de 100 marcos semánticos diferentes que incluye tanto verbos como adjetivos o sustantivos (y no todos los sustantivos son nominalizaciones). El FrameNet alemán contiene 648 unidades léxicas, que incluye verbos (493) y sustantivos (155), aunque como en el caso español, entre los sustantivos no solo se incluyen las nominalizaciones sino también otros tipos de sustantivos, como por ejemplo los nombres relacionales (que expresan partes del cuerpo o relaciones de parentesco)¹⁵.

OntoNotes

Otra propuesta que contempla las nominalizaciones deverbales es el proyecto OntoNotes¹⁶ (Hovy et al., 2006). El objetivo de este proyecto es desarrollar un corpus de un millón de palabras para cada una de las siguientes lenguas: inglés, árabe y chino. El proyecto consiste en anotar dicho corpus con los siguientes niveles de anotación: 1) anotación sintáctica, que sigue el mismo esquema de anotación propuesto para el inglés en el Penn TreeBank (Marcus et al., 1993); 2) anotación de la estructura argumental de los verbos, basándose en la propuesta de PropBank; 3) anotación de sentidos, se quiere anotar el sentido correspondiente de cada palabra tomando como referencia un conjunto de sentidos resultantes de la agrupación de *synsets* de WordNet; 4) anotación ontológica, se quiere asociar cada palabra a un nodo de la ontología Omega (Philpot et al., 2005); 5) anotación de la

¹²<http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index>

¹³<http://jfn.st.hc.keio.ac.jp/>

¹⁴<http://gemini.uab.es:9080/SFNsite>

¹⁵Sobre el FramNet japonés no hemos obtenido datos sobre las unidades léxicas que contiene.

¹⁶<http://www.bbn.com/ontonotes/>

correferencia, para cada palabra del corpus se quieren anotar las palabras correferentes; y, finalmente, 6) anotación de las entidades con nombre (*Named Entities*, en adelante NE). Este corpus tiene la particularidad que pretende anotarse con un 90 % de acuerdo entre los anotadores en cada uno de los niveles de anotación, lo que es un ambicioso objetivo. Con respecto a las nominalizaciones, en OntoNotes distinguen entre sentidos de nominalizaciones que realmente heredan el significado verbal ('construcción') y aquellos sentidos cuya denotación no se relaciona directamente con el significado del verbo ('consulta' como sinónimo de 'establecimiento'). Se puede llegar a deducir que el primer tipo se correspondería a los eventos y el segundo tipo a los resultados, aunque cabría esperar a ver ejemplos de anotación y comprobar dicha correspondencia.

WordNet¹⁷ (Fellbaum, 1998) es una base de datos léxica de grandes dimensiones (155.327 *synsets* en la versión inglesa), estructurada en forma de red semántica. En esta ontología los conceptos se definen como conjunto de sentidos sinónimos, *synsets*, relacionados mediante diferentes tipos de relaciones semánticas (hiponimia, meronimia, etc.). Algunos *synsets* proporcionan, además, una glosa en la que se incluye una definición corta y/o ejemplos. En este recurso no se especifica información sobre la estructura argumental, pero sí se contemplan distinciones similares a las de evento y resultado. WordNet, dada su rica (en opinión de muchos excesiva) granularidad de significados, normalmente incluye entre los sentidos correspondientes a las nominalizaciones deverbales uno que puede parafrasearse como "acción del verbo X" y otro que se parafrasea como "la cosa verbo-X-ada", que se corresponderían aproximadamente con las clases de evento y resultado respectivamente. Dado el éxito indiscutible del WordNet original de Princeton para el inglés, se han desarrollado extensiones para muchas otras lenguas. A día de hoy, existen 64 proyectos WordNets (Vossen and Fellbaum, 2009) para lenguas diferentes¹⁸ entre las que se incluyen el español (Atserias et al., 2004a) (véase la Tabla 2.3). Existen, además, proyectos que han agrupado en un único recurso los WordNets de distintas lenguas, estableciendo relaciones croslingüísticas. Es el caso de los recursos *EuroWordNet*¹⁹ (Vossen, 1998) y *Multilingual Central Repository*²⁰ (Atserias et al., 2004b). En el primero se integran los WordNets del holandés, italiano, español, alemán, francés, checo y estonio, y en el segundo, los WordNets del catalán, español, euskera, inglés e italiano. De esta manera, la información asociada a una de las lenguas es compartida por las demás lenguas.

A pesar de que el inglés es la lengua que cuenta con más recursos que representan las nominalizaciones deverbales, existen también recursos para otras lenguas más allá de los proyectos FrameNet y WordNet. Un ejemplo es el trabajo

¹⁷<http://wordnet.princeton.edu>

¹⁸http://www.globalwordnet.org/gwa/wordnet_table.htm

¹⁹<http://www.illc.uva.nl/EuroWordNet/>

²⁰<http://www.lsi.upc.edu/nlp/meaning/demo/demo.html>

The Essex Database of
Russian Verbs and their
Nominalizations

de Spencer and Zaretskaya (1999), que han creado de forma manual una base de datos, *The Essex Database of Russian Verbs and their Nominalizations*²¹ para el ruso que contiene cerca de 7.000 verbos y 5.000 nominalizaciones relacionadas. En esta base de datos se distingue entre aquellas nominalizaciones que nominalizan todo el evento y preservan la estructura argumental del verbo, de aquellas que denotan un resultado, concreto o abstracto, derivado de la acción del verbo, pero que no conservan la estructura argumental. Esta base de datos incluye información morfosintáctica y semántica sobre estos tipos de nominalizaciones. De hecho, a cada sentido nominal se le asigna una de las tres categorías de sustantivos propuestas por Grimshaw (1990), es decir, evento complejo, evento simple o resultado.

NOMAGE

En el marco del proyecto Nomage (Balvet et al., 2010, 2011)²², que se centra en la descripción de las propiedades aspectuales de las nominalizaciones del francés, se ha realizado la anotación de los sustantivos deverbales del corpus FrenchTreeBank (Abeillé et al., 2000) (4.042 ocurrencias de sustantivos deverbales en total). Estas ocurrencias fueron anotadas de acuerdo a una tipología de tres clases aspectuales generales (evento, estado, objeto) por anotadores no especializados que aplicaban una serie de pruebas de combinación (comprobaban que la nominalización combinara bien o mal con tipos de determinantes como *plusieurs*, ‘varios’, construcciones verbales como *avoir lieu*, ‘tener lugar’, etc.) para la clasificación de las nominalizaciones deverbales. A partir de esta anotación, se ha desarrollado de manera manual un léxico de 746 entradas léxicas, correspondientes a los lemas de las ocurrencias del corpus previamente anotadas. En el léxico, además de la definición de la nominalización, se recoge también información sobre el verbo base, la estructura argumental, el tipo aspectual de la nominalización y los ejemplos del corpus asociados a dicha entrada léxica. Sin embargo, cabe señalar que este léxico se ha desarrollado de manera manual por anotadores especializados y las clases aspectuales con los que se ha asociado a los lemas nominales no son las tres clases generales del corpus sino las once clases más específicas (véase la Sección 2.1.1). A pesar de todo, parece existir un alto grado de correspondencia entre las clases generales asociadas por anotadores no expertos y las once más específicas asignadas por los anotadores expertos.

Copenhaguen
Dependency Treebank,
CDT

Por último, las nominalizaciones están siendo anotadas de manera manual en el *Copenhaguen Dependency Treebank* (Hoeg Muller, 2010) (CDT, en adelante), un proyecto cuyo objetivo es la creación de un corpus de dependencias paralelo para el danés, inglés, alemán, italiano y español de 80.000 palabras para cada lengua. En el nivel del SN se establecen dos tipos de dependencias: 1) la sintáctica, que indica el tipo de función sintáctica (objeto, sujeto, atributiva) que le correspon-

²¹http://privatewww.essex.ac.uk/~spena/res_interests.htm

²²<http://stl.recherche.univ-lille3.fr/programmesetcontrats/NOMAGE/NOMAGEenglish.html>

2. NOMINALIZACIONES DEVERBALES: ESTADO DE LA CUESTIÓN

de al complemento de la nominalización, y 2) la semántica, que establece una serie de relaciones semánticas entre la nominalización y sus argumentos-complementos (agente, paciente, experimentador, recipiente y lugar).

A continuación presentamos una tabla resumen de los diferentes recursos presentados (Tabla 2.3). La primera columna recoge el nombre del recurso; en la segunda se detalla qué tipo de recurso es, léxico, corpus, etc.²³; la tercera nos informa del tamaño del recurso en número de entradas léxicas, *synsets* o palabras según el tipo de recurso; la cuarta nos indica si el recurso se ha creado de manera automática (A) o manual (M); y finalmente, la quinta y la sexta columna indican si en los recursos correspondientes se representa el tipo denotativo y la estructura argumental (EA) respectivamente.

Recursos	Tipo	EL/Palabras	A o M	Denotación	EA
NomLex:Inglés		1.025 EL	M	-	+
NomLex-Plus:Inglés		7.050 EL	A	-	+
FrameNet:Inglés		11.600 EL	M	+	+
FrameNet:Español	Léxicos	1.200 EL	M	+	+
FrameNet:Alemán		648 EL	M	+	+
FrameNet:Japonés		? EL	M	+	+
Nomage-Francés		815 EL	M	+	+
WordNet:Inglés	Ontologías	155.327 syn.	M	+	-
WordNet:Español		67.351 syn.	M	+	-
Essex-Data-Base:Ruso	Base de datos	800 EL	M	+	+
CDT		80.000 pal.	M	-	+
Nomage	Corpus	1 millón pal.	M	+	+
NomBank		4,5 millones pal.	M	-	+
OntoNotes		1 millón pal.	M	+	-

Tabla 2.3: Recursos lingüísticos que representan las nominalizaciones deverbales

La Tabla 2.3 pone de manifiesto que salvo en el proyecto FrameNet, no existe ningún recurso para el español que represente las nominalizaciones deverbales. En este sentido, y dado que el FrameNet español solo tiene 1.200 unidades léxicas entre sustantivos, verbos y adjetivos, creemos que los recursos desarrollados en este trabajo, AnCora-Nom (Peris and Taulé, 2011a), un léxico de nominalizaciones

²³Cabe mencionar que entre las diferentes maneras de representar las nominalizaciones la terminología es diversa y no siempre existe una delimitación clara entre las distintas maneras. Por ejemplo, el proyecto FrameNet se define como una base de datos léxica, esto es, un léxico, pero al mismo tiempo se podría considerar una ontología por las relaciones que se establecen entre los distintos *frames* o un corpus, si se tiene en cuenta que para cada *frame* se especifican los ejemplos de los que se obtiene la evidencia empírica.

deverbales del español, y el corpus AnCora-Es enriquecido con la anotación de las nominalizaciones de verbales del español (Peris et al., 2010b), suponen una importante contribución al desarrollo de la tecnología lingüística del español, además de constituir dos fuentes de información valiosas para el análisis lingüístico.

2.2.2. Sistemas

En esta sección presentamos los sistemas automáticos que se han desarrollado para el tratamiento computacional de las nominalizaciones de verbales desde una perspectiva semántica. Los trabajos que reseñaremos se centran básicamente en la detección de relaciones semánticas y están mayoritariamente desarrollados para el inglés. Sin embargo, existen dos líneas de trabajo claramente diferenciadas. La primera se centra en la detección de relaciones semánticas del tipo causa-efecto, parte-todo, contenedor-contenido, etc. Estas relaciones pueden establecerse entre: a) pares de sustantivos que pertenecen a distintos SNs, la Tarea 4 del SemEval 2007 (Girju et al., 2009) y la Tarea 8 del SemEval 2010 (Hendrickx et al., 2009, 2010) o b) sustantivos que forman parte de lo que en inglés se conoce como *compound nouns*, ‘compuesto nominal’ (N+N), por ejemplo *colon cancer*, ‘cáncer de colon’ (Moldovan et al., 2004; Girju et al., 2004, 2005). Existe una variación de esta última tarea que consiste en detectar la relación entre los dos sustantivos del compuesto nominal mediante la paráfrasis formada por un verbo y una preposición, objetivo de la Tarea 9 de SemEval 2010 (Butnariu et al., 2009, 2010; Nakov, 2007). El problema de estos sistemas respecto a nuestro trabajo es que, a pesar de que incluyen las nominalizaciones, no están estrictamente centrados en ellas sino en todo tipo de sustantivos.

La segunda línea de trabajos se centra en la asignación de argumentos y papeles temáticos a los complementos de la nominalización. En este segundo grupo hay trabajos que focalizan en la detección de argumentos dentro del SN como son los de Lapata (2002); Hull and Gomez (2000); Gurevich and Waterman (2009); Padó et al. (2008) y la CoNLL-2008 Shared Task on *Joint Parsing of Syntactic and Semantic Dependencies* (Surdeanu et al., 2008)²⁴, y otros que se centran en la detección de los argumentos fuera del SN (Gerber et al., 2009; Gerber and Chai, 2010) y la Tarea 10 de SemEval 2010 (Ruppenhofer et al., 2009, 2010). Es esta segunda línea la que más nos interesa puesto que está estrechamente vinculada con nuestro trabajo, especialmente con el sistema desarrollado (RHN) para la anotación de los argumentos de las nominalizaciones de verbales en el corpus AnCora-Es.

A pesar de que la mayoría de estos trabajos reconocen la distinción entre nominalizaciones eventivas y resultativas, los sistemas desarrollados no tienen como

²⁴<http://www.clips.ua.ac.be/conll2008/>

objetivo distinguir entre ambas denotaciones. De hecho, sistemas automáticos para la desambiguación entre sentidos nominales eventivos y resultativos solo conocemos el desarrollado por Eberle et al. (2011) que se centra en la desambiguación de las nominalizaciones en *-ung* del alemán. Un trabajo relacionado también es el de Creswell et al. (2006) que presentan un clasificador entre sentidos nominales eventivos y no-eventivos para el inglés. En esta sección nos detendremos primero en estos dos trabajos, a continuación reseñaremos brevemente los sistemas de detección de relaciones semánticas entre el núcleo y los componentes del SN y finalmente, cerraremos la sección y el capítulo, con los sistemas que asignan automáticamente los argumentos de la nominalización de verbal.

2.2.2.1. Sistemas que tratan la distinción entre evento y resultado

La noción de evento, pero no la de resultado, está presente en el trabajo de Creswell et al. (2006). En este trabajo se presenta un sistema automático que distingue entre sustantivos que denotan eventos y sustantivos que denotan no-eventos para el inglés. A partir de dos listas de sustantivos no ambiguos compilados manualmente, una de sustantivos eventivos (en total 95) y otra de sustantivos no eventivos (295), y de un corpus formado por 170.000 documentos previamente analizado sintácticamente con un analizador de dependencias, extraen información sobre los sustantivos y el contexto de cada una de las dos clases de sustantivos. Los atributos que extraen son tuplas del tipo <sustantivo, relación sintáctica > o <relación sintáctica, sustantivo>. Con esta información, desarrollan un método probabilístico que ante una ocurrencia de un sustantivo la clasifica en una de las dos clases. El método se basa en la construcción de dos modelos bayesianos generativos, uno para generar sustantivos eventivos y otro sustantivos no eventivos. Los dos modelos responden a una distribución multinomial sobre los diferentes atributos. El clasificador resulta de la comparación del resultado producido por los dos modelos. Consiguen, con información de la palabra y el contexto, un 64,5 % de corrección, que asciende a 79,5 % mediante el uso de técnicas de *bootstrapping*, aumentando en sucesivas iteraciones los vocabularios iniciales con los sustantivos mejor puntuados en los modelos generativos, pudiendo el sustantivo pertenecer a cualquiera de las listas iniciales o a ninguna de ellas. Sin embargo, una diferencia básica entre este trabajo y el nuestro, es que ellos no se centran solo sobre nominalizaciones de verbales sino sobre toda clases de sustantivos, por lo que la distinción entre evento y no-evento, que afecta a toda clase de sustantivos, no es comparable a la distinción entre evento y resultado de las nominalizaciones de verbales. Como ejemplo, considérese que como palabras de lista de sustantivos no-eventos encontramos sustantivos como *airport*, ‘aeropuerto’, o *electronics*, ‘electrónica’ lo que demuestra que el tipo de distinción y la información utilizada para establecerla no es comparable.

Creswell et al., 2006

En el trabajo de Eberle et al. (2011) se mantiene que las nominalizaciones de verbales del alemán en *-ung*, el prefijo nominalizador más productivo de esta lengua comparable a nuestro sufijo *-ción*, pueden denotar un evento, un estado y un objeto-resultado. Sin embargo, no siempre estas nominalizaciones son triplemente ambiguas sino que según la clase semántica del verbo base la nominalización podrá tener tres, dos o solo una de las tres denotaciones posibles. En concreto, el estudio se centra en aquellas nominalizaciones en *-ung* que derivan de verbos de dicción ('decir', 'declarar', 'comentar', 'explicar') y que aparecen incrustadas en sintagmas preposicionales (SP) introducidos por la preposición *nach*, 'hacia'. Según los autores, este tipo concreto de nominalización puede denotar o bien un evento o bien una proposición, que es un tipo de objeto específico relacionado con los verbos de dicción. Eberle et al. (2008) presentan un sistema que clasifica este tipo de denotaciones en base a nueve criterios, denominados indicadores. El sistema genera una representación semántica de las oraciones en forma de *FU-DRS -flat underspecified discourse representation structures-* (Eberle, 2004) de la cual extrae los criterios para la clasificación de la nominalización según la denotación. A partir de estos criterios, el sistema calcula la denotación preferida para la nominalización en función de los pesos asignados a cada criterio de manera preestablecida. Esta herramienta se ha aplicado a 100 oraciones en las que los criterios son accesibles al sistema y la corrección lograda es del 82 %. Si bien este trabajo no es directamente comparable con el ADN-Classifer ya que nosotros trabajamos con una gama más amplia de sufijos y no limitamos el tipo de verbo base de la nominalización, sí es cierto que es el que guarda una relación más estrecha con el clasificador de denotaciones que se ha construido en el marco de este trabajo. En la Sección 7.4 presentamos una comparación parcial de nuestro clasificador con este.

2.2.2.2. Sistemas de detección de relaciones semánticas entre pares de sustantivos

Como se ha visto anteriormente, en tareas de diferentes ediciones de SemEval se han presentado trabajos que tratan de detectar las relaciones semánticas existentes entre dos sustantivos, que bien pertenecen a SNs distintos o bien forman parte del mismo SN (*compound noun*). Sin embargo, la mayoría de estos trabajos tienen la particularidad que no se centran en nominalizaciones de verbales sino que pueden ser núcleo del SN todo tipo de sustantivos, como ocurre por ejemplo en el trabajo de Moldovan et al. (2004) o en la Tarea 4 de la competición SemEval 2007 (Girju et al., 2009)²⁵. Aquí solo nos referiremos a los trabajos que implican

²⁵Inicialmente, los sistemas trataban de extraer relaciones simples entre entidades con nombre. En esta tarea, los sistemas basados en realimentación (*bootstrapping*) obtuvieron buenos resultados. A partir de SemEval 2007, se extiende la tarea a la extracción de relaciones entre menciones

únicamente a las nominalizaciones deverbales.

En el trabajo de Girju et al. (2004) se clasifican relaciones semánticas que se dan entre el sustantivo núcleo y el modificador de los SNs en inglés. Se distinguen cinco patrones sintácticos en los que bien el sustantivo núcleo o bien el sustantivo base del modificador son una nominalización de verbal. Se distingue entre 35 posibles relaciones semánticas, como por ejemplo agente, temporal, parte-todo, causa, frecuencia, si bien parece que cuando el núcleo del SN es una nominalización de verbal la relación que se da es una de tipo predicado-argumento. Con un algoritmo de aprendizaje basado en *Support Vector Machine* (SVM) consiguen un 72 % de corrección para las construcciones Nombre+Nombre, un 67 % de corrección para las construcciones Nombre+Genitivo'S, un 61 % de corrección para las construcciones Nombre+Genitivo-of, un 64 % de corrección para las construcciones Nombre+SP y un 74 % de corrección para las construcciones Nombre+cláusulas de relativo.

Girju et al., 2004

2.2.2.3. Sistemas de detección de argumentos de las nominalizaciones

Existen diferentes trabajos que se centran en la anotación de argumentos de las nominalizaciones deverbales basándose sobre todo en información verbal. Es decir, todas las propuestas que a continuación describimos asumen que la estructura argumental de las nominalizaciones deriva de los verbos base correspondientes, si bien la manera en qué se anotan estos argumentos y las técnicas utilizadas son diferentes: métodos probabilísticos (Lapata, 2002; Gurevich and Waterman, 2009), reglas heurísticas (Hull and Gomez, 2000; Gurevich et al., 2006), aprendizaje automático no supervisado (Padó et al., 2008) y supervisado (Surdeanu et al., 2008). Tampoco hay unanimidad en el tipo de argumento anotado: en los trabajos de Lapata (2002) y Gurevich and Waterman (2009) se anotan los argumentos de las nominalizaciones con etiquetas más sintácticas, aquellos que se corresponderían con el sujeto verbal (+subj) y aquellos que lo harían con el objeto verbal (+obj); en cambio, en Padó et al. (2008) y Surdeanu et al. (2008) se utilizan etiquetas semánticas, de FrameNet en el primer caso y de NomBank en el segundo. Entre todos estos sistemas, nos interesan especialmente aquellos que parten de información verbal para la anotación de los argumentos de las nominalizaciones puesto que siguen la misma hipótesis que nuestro trabajo: a partir de la información verbal se pueden inferir los argumentos de las correspondientes nominalizaciones deverbales. En este sentido, se excluyen, por lo tanto, los sistemas supervisados de etiquetado semántico nominal ya que aprenden a partir de información nominal

nominales dominadas por un nombre común y se amplía el rango de relaciones a extraer a relaciones más complejas (Girju las denomina “relaciones contingentes”) como la causalidad, la instrumentación o formas de meronimia.

previamente anotada en corpus (Surdeanu et al., 2008) y no utilizan información verbal para anotar los argumentos de las nominalizaciones.

Hull and Gomez, 2000

Una de las primeras propuestas para la anotación de la estructura argumental de las nominalizaciones deverbales a partir de información verbal es la de Hull and Gomez (2000). Según este enfoque, para determinar la interpretación semántica de las nominalizaciones, además de saber el significado de la nominalización, es también necesario otorgar un significado a los complementos nominales (de hecho, a veces no se puede obtener el significado de la nominalización si no se interpretan primero sus complementos). Los autores parten de una base de conocimiento verbal en la que se especifican los sentidos verbales y sus correspondientes restricciones de subcategorización y mantienen que para anotar la estructura argumental de las nominalizaciones tan solo es necesario especificar las restricciones propias de la nominalización (por ejemplo, preposición regida diferente que la correspondiente verbal, orden específico de los argumentos, restricciones sobre la realización de argumentos por constituyentes, entre otros). Esto lo hacen para un grupo de diez nominalizaciones: *arrest*, ‘arresto’; *birth*, ‘nacimiento’; *capture*, ‘captura’; *control*, ‘control’; *defense*, ‘defensa’; *execution*, ‘ejecución’; *murder*, ‘asesinato’; *nomination*, ‘nominación’; *publication*, ‘publicación’; y *trade*, ‘comercio’. A partir de aquí diseñan tres algoritmos: el primero tiene como objetivo determinar el sentido verbal concreto del que deriva la nominalización y, por lo tanto, identificar qué roles semánticos deben satisfacer los complementos nominales; el segundo trata de identificar qué complementos de la nominalización satisfacen algún rol semántico, primero empezando por los SPs puesto que son más fáciles de identificar y así se descartan roles semánticos para el resto de complementos de la nominalización (adjetivos y genitivos); el tercer y último algoritmo tiene como objetivo determinar el *concepto* verbal de la nominalización, si aún no se conoce, y reevaluar cada complemento de la nominalización para asegurar que se ha encontrado un rol semántico adecuado. Aplican estos tres algoritmos a 1.247 ocurrencias de las diez nominalizaciones seleccionadas y consiguen muy buenos resultados en la interpretación de los complementos genitivos (93 % de corrección), de los SPs (96 %) y de los SAs (71 %). Sin embargo, estos resultados son dudosamente extrapolables porque dependen de unas reglas/restricciones especificadas manualmente para estas diez nominalizaciones.

Lapata, 2002

Una aproximación más próxima a nuestros intereses es la de Lapata (2002). En este trabajo se estudian las construcciones de los SNs del inglés formadas por dos sustantivos (N+N) en la que el núcleo es la nominalización deverbal. Lapata enfoca el problema desde la ambigüedad del sustantivo modificador, que según la autora se puede interpretar como el sujeto (+subj), el objeto (+obj) o como un complemento preposicional del verbo base correspondiente. En este trabajo se trata de desambiguar entre la interpretación de (+subj) o (+obj) de los sustantivos modificadores de las nominalizaciones deverbales. Para ello, se establece

que un sustantivo modificador tendrá más probabilidad de ser (+subj) u (+obj) de una nominalización en función de si ese sustantivo modificador es más frecuente como objeto o sujeto del verbo base de la nominalización. Para calcular estas probabilidades Lapata extrae tuplas de <V+N-obj> (615.328 en total) y tuplas <V+N-subj> (588.333 en total) del *British National Corpus* (BNC) (Aston and Burnard, 1998). Aplica esta función de probabilidad a 796 nominalizaciones que cumplen un requisito: tienen como complemento un sustantivo que solo puede tener la interpretación de (+subj) o (+obj). Dado que no todos los sustantivos modificadores aparecen en las tuplas extraídas, tienen que aplicarse técnicas de suavizado (*smoothing*) para hacer frente a los casos infrarrepresentados (*data sparseness*). El mejor resultado sin tener en cuenta el sufijo específico de la nominalización es de 75,8 % de corrección y de 76,3 % si el sufijo sí se tiene en cuenta. La pequeña mejoría (0,5 %) se explica porque sufijos como *-er* en inglés indican que la nominalización es agentiva, esto es, incorpora el sujeto por lo que el sustantivo modificador solo puede ser objeto. En este trabajo también se experimenta con diferentes técnicas de suavizado, con la inclusión de contexto y con la combinación de ambas cosas. El uso de las técnicas de suavizado permite aumentar la corrección hasta un 80,4 %. El contexto se incluye ampliando la ventana del N(sustantivo modificador)+N(nominalización) a diferentes lemas tanto por la derecha como por la izquierda y usando tanto la información del lema como la etiqueta de *Part of Speech*, en adelante PoS. Aunque experimentan con diferentes ventanas de contexto y los dos tipos de información, el mejor resultado (68,6 % de corrección) se consigue con información de lemas con la ventana abierta en dos lemas por la derecha. La combinación de técnicas de suavizado y la inclusión de contexto logra un 85,1 % de corrección.

También para el inglés, el trabajo de Gurevich and Waterman (2009) asigna las etiquetas sintácticas (+Subj) y (+Obj) a los complementos de las nominalizaciones deverbales, aunque este trabajo está centrado en nominalizaciones derivadas de verbos transitivos y solo anota con estas etiquetas los SPs introducidos por la preposición *of*, ‘de’ y los determinantes posesivos. Los autores presentan tres modelos diferentes el objetivo de los cuales es mejorar el sistema para la anotación de las nominalizaciones deverbales. El sistema de anotación consiste en un grupo de reglas heurísticas similares a las descritas en Gurevich et al. (2006). Estas heurísticas se resumen de la siguiente manera: los argumentos de las nominalizaciones agentivas (‘diseñador’) son +Obj, los de las nominalizaciones de paciente (‘traducción’) son +Subj y en las nominalizaciones eventivas (‘creación’) los determinantes posesivos son +Subj y los SP en *of*, ‘de’ +Obj. Estas heurísticas, sin embargo, no siempre se manifiestan adecuadas para la asignación de las etiquetas sintácticas, por lo que se proponen tres modelos nuevos que siguen una intuición similar a la propuesta por Lapata (2002): si un argumento X es preferido como sujeto o como objeto de un verbo, entonces será preferido como tal si complementa

Gurevich and Waterman,
2009

a la nominalización correspondiente a aquel verbo. Para examinar esta intuición se extraen todas las parejas verbo-argumento y nominalización-argumento de la Wikipedia en inglés analizadas sintácticamente; para las primeras se tiene en cuenta la relación entre verbo y argumento (+Subj, +Obj) y para las segundas el tipo de argumento (posesivo, SP, etc.). A partir de aquí se desarrollan tres modelos: el primero y más simple, compara el número de argumentos de la nominalización que muestran una preferencia +Subj con aquellos que muestran una preferencia más +Obj (a partir de la comparación con las parejas verbo-argumento correspondientes). Si alguna de las dos preferencias es 1,5 veces mayor que la otra, entonces se le asigna ese rol; el segundo modelo incorpora además el rasgo de la animación del complemento (es decir, si es animado o no-animado) y el tercero especifica preferencias léxicas de los roles semánticos de las nominalizaciones deverbales (es decir, si un determinado rol semántico tiende a ser realizado mediante un SP con una preposición específica). El mejor resultado se consigue con este último modelo, que logra un 82 % de corrección en la anotación de los SP en *of*, *'de'*, como argumentos de las nominalizaciones y un 85 % en los determinantes posesivos.

El problema de los sistemas hasta ahora descritos es que se centran principalmente en un número escaso de etiquetas, dos en concreto, y no tienen en cuenta todos los posibles argumentos que pueden tener las nominalizaciones deverbales. Sin embargo, los sistemas de SRL para sustantivos, desarrolladas básicamente para el inglés, anotan una gama más amplia de argumentos. Estos sistemas se basan en técnicas de aprendizaje automático. Entre ellos distinguimos dos aproximaciones: el aprendizaje automático no supervisado (Padó et al., 2008) y el aprendizaje automático supervisado (Che et al., 2008), (Johansson and Nugues, 2008), (Zhao and Kit, 2008) y (Ciaramita et al., 2008), sistemas presentados en la CoNLL-2008 Shared Task on *Joint Parsing of Syntactic and Semantic Dependencies* (Surdeanu et al., 2008)²⁶. Como avanzamos al inicio de esta sección, nos centramos en los sistemas de SRL nominal no supervisado puesto que los sistemas supervisados no parten de información verbal sino de información nominal previamente anotada.

En el trabajo de Padó et al. (2008) se aborda la tarea de SRL partiendo únicamente de información verbal, concretamente, usan la información relativa a los roles semánticos de los verbos representados en FrameNet para asignar roles semánticos a las nominalizaciones deverbales correspondientes. A partir de una lista de 265 parejas verbo-nominalización obtenida de FrameNet 1.3, utilizan 26.479 instancias verbales como datos para el aprendizaje y 6.502 ocurrencias nominales como datos de evaluación de los diferentes modelos. En la tarea de SRL se distinguen dos subtareas, la de reconocimiento de argumentos y la de asignación de argumentos/roles semánticos. En este trabajo la primera de ellas sigue una

Padó et al., 2008

²⁶<http://www.clips.ua.ac.be/conll2008/>

regla bastante simple: todos los constituyentes del SN cuyo núcleo es la nominalización deverbal son considerados como potenciales argumentos. A pesar de la simplicidad, dado que no intentan discriminar entre argumentos obligatorios y adjuntos, consiguen una F1 (definida como la media armónica de precisión y cobertura) de 82,83 % si solo se tienen en cuenta los constituyentes dentro del SN y una F1 de 76,89 % si se tienen en cuenta los constituyentes de dentro y fuera del SN. Más interesante es la tarea de asignación de argumentos en la que presentan tres clases de modelos: (i) el modelo simple basado en atributos léxico-semánticos, (ii) el modelo simple basado en atributos estrictamente sintácticos y (iii) dos modelos distribucionales que calculan la etiqueta semántica del argumento a partir de medidas de similitud semántica entre los argumentos de la nominalización y del verbo correspondiente, teniendo en cuenta o bien el lexema del argumento de la nominalización o bien su función sintáctica. Se consideran modelos distribucionales porque miden la similitud semántica por la distancia entre representaciones vectoriales de lexemas en un espacio de coocurrencia semántica.

A partir de estos modelos, experimentan con modelos híbridos, que combinan los modelos simples con los distribucionales, y el mejor resultado (56,42 % de corrección) se consigue con un modelo que combina atributos sintácticos con medidas de similitud semántica basadas en la función sintáctica del argumento de la nominalización y el argumento verbal. El resto de modelos híbridos logran alrededor de un 50 % de corrección. En solitario, es decir, sin combinarse con otros modelos, solo los modelos distribucionales superan el caso base (43 % de corrección): el modelo que calcula la similitud semántica en base al lexema del argumento nominal logra un 44,5 % de corrección y el modelo que calcula la similitud semántica en base a la función sintáctica del argumento nominal obtiene un 52 % de corrección.

Las aproximaciones supervisadas para el SRL nominal parten de información nominal previamente anotada en corpus, por lo que el resultado es mejor que en los métodos no supervisados. Por ejemplo, en la CoNLL-2008 Shared Task on *Joint Parsing of Syntactic and Semantic Dependencies* (Surdeanu et al., 2008), el sistema que logra un mejor resultado (una F1 de 76,64 %) es el de Che et al. (2008). Sin embargo, los resultados no son comparables ya que la calidad de la información de la que parte el aprendizaje y el coste de obtenerla son muy diferentes.

Che et al., 2008

Relacionado con estos sistemas, tenemos los sistemas que tienen en cuenta los argumentos implícitos de las nominalizaciones deverbales. Palmer et al. (1986) propusieron uno de los primeros métodos automáticos para recuperar argumentos extra oracionales. Su aproximación consiste en detectar los argumentos implícitos mediante el uso de conocimiento sobre ciertos predicados y sobre ciertas tendencias de cadenas de correferencia en oraciones pertenecientes a un mismo dominio temático. Sin embargo, este método se aplica a un dominio específico (informes

Palmer et al., 1986

Ruppenhofer et al., 2009, 2010

de mantenimiento de equipos informáticos) por lo que resulta difícil imaginar la implementación de este método para dominios no restringidos. Más recientemente, en la Tarea 10 de SemEval 2010 (Ruppenhofer et al., 2009, 2010) se evaluaron distintos sistemas encargados de identificar los argumentos implícitos, siguiendo la tipología propuesta en Fillmore and Baker (2001), de varios tipos de predicados (verbos, sustantivos, adjetivos y preposiciones). Los organizadores proporcionaban un corpus formado por textos literarios de ficción y etiquetado con argumentos explícitos e implícitos siguiendo el esquema de anotación de FrameNet que constaba de 438 oraciones con 1.370 predicados para el entrenamiento y 525 oraciones con 1.703 predicados para el test. Solo tres sistemas se presentaron a dicha tarea, y teniendo en cuenta que los equipos podían elegir entre realizar SRL estándar, es decir, anotar solo los argumentos explícitos, o bien detectar los argumentos implícitos o ambas cosas, es decir, anotar tanto argumentos explícitos como implícitos, solo dos optaron por la detección de argumentos implícitos, consiguiendo un 63,4 % (Semafor Sytem) y 8 % (GETARUNS++) de F1 respectivamente.

Gerber and Chai, 2010

Sin embargo, dado que no ofrecen resultados por tipo de predicados tratados, nos parece más interesante para nuestra investigación el trabajo de Gerber and Chai (2010) que se centra en los argumentos implícitos de las nominalizaciones deverbales. Estos autores ya habían previamente demostrado la importancia de tener en cuenta los argumentos implícitos de los predicados nominales o bien de dejar fuera los sustantivos con argumentos implícitos en los sistemas de SRL nominal. Los autores argumentan que, de lo contrario, las muestras de aprendizaje para llevar a cabo SRL nominal no son suficientes (porque hay sustantivos que tienen argumentos implícitos) para dar lugar a modelos adecuados de SRL nominal (Gerber et al., 2009). En el trabajo de Gerber and Chai (2010) se seleccionan los 10 nombres más frecuentes²⁷ con sentidos no ambiguos del Penn TreeBank (Marcus et al., 1993) y se anotan manualmente los argumentos implícitos nucleares de estos. A partir de esta anotación (1.253 ocurrencias en total), se separan dos corpus, el de entrenamiento (816 ocurrencias) y el de test (437 ocurrencias). Para la detección de los argumentos implícitos que se tienen que anotar, se consideran aquellos argumentos que no están anotados en la ocurrencia de NomBank pero sí se encuentran como posibles argumentos de la nominalización en el léxico asociado a NomBank, es decir, NomLex-Plus. Los candidatos a satisfacer esos potenciales argumentos implícitos son los constituyentes anotados como argumentos de un verbo en PropBank o una nominalización en NomBank. A partir del corpus de entrenamiento, se aplica un modelo de regresión lógica basado en rasgos (un total de 14) que consigue un resultado promedio para los diez nombres de 42,3 %

²⁷*price*, ‘precio’; *sale*, ‘venta’; *investor*, ‘inversor’; *fund*, ‘fund’; *loss*, ‘pérdida’; *plan*, ‘plan’; *investment*, ‘inversión’; *cost*, ‘coste’; *bid*, ‘bid’; y *loan*, ‘préstamo’.

de F1, siendo el mejor resultado individual un 83,3 % y el peor un 15,4 %. A pesar de que los resultados no son espectaculares, este trabajo abre una nueva línea de investigación que permite detectar los argumentos implícitos nominales, que según Gerber and Chai suponen el 65 % de los argumentos nominales.

A continuación presentamos una tabla resumen de los diferentes sistemas presentados que tratan específicamente las nominalizaciones deverbales (Tabla 2.4). La primera columna identifica los diferentes sistemas; la segunda indica la lengua para la que trabajan dichos sistemas, en la tercera se especifica el objetivo, la tarea a desarrollar por el sistema, y en la cuarta se indica el enfoque técnico seguido.

Como se puede ver en la Tabla 2.4, no existe ningún sistema automático diseñado para el tratamiento computacional de las nominalizaciones deverbales en español. Nuestro trabajo quiere suplir dicha carencia puesto que hemos diseñado dos herramientas que tratan computacionalmente las nominalizaciones deverbales del español. La primera anota automáticamente los argumentos explícitos de las nominalizaciones deverbales (RHN) y la segunda se centra en la desambiguación del tipo denotativo de las nominalizaciones deverbales (ADN).

Sistemas	Lengua	Tarea	Enfoque Técnico
Eberle et al. (2011)	Alemán	Desambiguación entre denotaciones	Reglas heurísticas + peso
Creswell et al. (2006)	Inglés	Desambiguación entre eventos y no-eventos	Método Probabilístico
Gerber and Chai (2010)	Inglés	Anotación de argumentos implícitos	Aprendizaje automático supervisado
Girju et al. (2004)	Inglés	Detección de relaciones en SNS	SVM
Gurevich et al. (2006)	Inglés	Anotación parcial de argumentos explícitos	Reglas heurísticas
Gurevich and Waterman (2009)	Inglés	Anotación parcial de argumentos explícitos	Método Probabilístico
Hull and Gomez (2000)	Inglés	Anotación de argumentos explícitos	Reglas heurísticas
Lapata (2002)	Inglés	Anotación parcial de argumentos explícitos	Método Probabilístico
Padó et al. (2008)	Inglés	Anotación de argumentos explícitos	Aprendizaje automático no supervisado
Surdanu et al. (2008)	Inglés	Anotación de argumentos explícitos	Aprendizaje automático supervisado

Tabla 2.4: Sistemas automáticos para el tratamiento computacional de las nominalizaciones deverbales

Parte II

Estructura Argumental

CAPÍTULO 3

ESTRUCTURA ARGUMENTAL DE LAS NOMINALIZACIONES DEVERBALES: ESTUDIO EMPÍRICO

En este capítulo se presenta la parte del estudio lingüístico basado en corpus dedicada a la estructura argumental de las nominalizaciones deverbales. El estudio se llevó a cabo como primera aproximación a las nominalizaciones deverbales del español y se centró en los dos fenómenos lingüísticos que nos interesaban de las mismas: la diferencia denotativa entre evento y resultado, presentada en el Capítulo 5, y la estructura argumental, objetivo de este capítulo.

La hipótesis de partida asumida en este trabajo es que las nominalizaciones deverbales heredan la estructura argumental de los verbos de los que derivan morfológicamente o se relacionan semánticamente (véase la Sección 1.1.1), pero nos interesaba saber en qué medida y cómo. Concretamente, qué tipo de argumentos tienen las nominalizaciones, en qué posición se realizan y cómo, es decir mediante qué constituyentes y en qué orden. En definitiva, nuestro objetivo radica en analizar la estructura interna de las nominalizaciones deverbales. El análisis consiste en observar estos hechos y para llevarlo a cabo se marcaron los constituyentes que podían considerarse argumentos (de la misma manera que en los verbos) de 817 sustantivos deverbales (que corresponden a un total de 3.077 ocurrencias) del corpus AnCora-Es (Taulé et al., 2008; Recasens and Martí, 2010). Antes de adentrarnos en el análisis lingüístico propiamente dicho (Sección 3.3), describimos cómo se ha obtenido la muestra de datos (Sección 3.1) y el esquema de anotación utilizado (Sección 3.2). Para terminar el capítulo, presentamos unas conclusiones (Sección 3.4).

3.1. Extracción de la muestra de datos

Corpus

La muestra de datos analizada consta de 817 sustantivos deverbales, correspondientes a 3.077 ocurrencias, el total de sustantivos deverbales que aparecen en un subconjunto de 100.000 palabras del corpus AnCora-Es. Este subconjunto está formado por 75.000 palabras de Lexesp (Sebastián et al., 2000), un corpus equilibrado de 6 millones de palabras, y por 25.000 palabras extraídas de la agencia española de noticias EFE¹. AnCora-Es es un corpus del español de 500.000 palabras que se constituye básicamente de textos periodísticos² anotados a diferentes niveles lingüísticos: morfología (PoS y lemas), sintaxis (constituyentes y funciones sintácticas), semántica (estructura argumental de los verbos, papeles temáticos, clases semánticas verbales, entidades nombradas y sentidos nominales de WordNet) y pragmática (correferencia)³.

El proceso de extracción de los datos fue llevado a cabo semiautomáticamente en dos etapas: 1) la extracción automática de sustantivos y 2) la selección manual de las nominalizaciones deverbales.

Extracción automática

Para llevar a cabo la extracción automática se partió de una lista predefinida de 13 sufijos (*-a*, *-aje*, *-azo*, *-ión/-ción/-sión/-ón*, *-deral/-era*, *-dal/-do*, *-dura/-ura*, *-e*, *-era*, *-ido*, *-miento/-mento*, *-ncia/-nza*, *-o/-eo*) que según Santiago and Bustos (1999) pueden dar lugar a nombres de acción o resultado (recuérdese que esta misma muestra de datos es utilizada para el estudio empírico de la denotación) y que toman verbos como base del proceso de derivación⁴. Sobre el subconjunto de 100.000 palabras de AnCora-Es se extrajeron automáticamente aquellos nombres comunes (NC) etiquetados en el corpus⁵ cuya terminación coincidía con estos 13 sufijos y sus correspondientes alomorfos (22 terminaciones en total). Como resultado se obtuvieron un total de 4.516 lemas nominales distintos.

Selección manual

Tras la extracción automática fue necesaria la selección manual de todos aquellos nombres claramente deverbales y con un significado de acción y/o resultado. Se descartaron aquellos nombres cuyas terminaciones coincidían con las formas sufijales mencionadas pero que eran en realidad parte de la raíz nominal, como ocurre por ejemplo con el sustantivo ‘avión’. También se excluyeron los sustantivos derivados de categorías morfosintácticas que no fueran verbos y que no

¹Este subconjunto de 100.000 palabras forman el corpus 3LB (Civit and Martí, 2004), que más tarde ha sido parte del corpus AnCora-Es.

²De las 500.000 palabras de AnCora-Es, 225.000 provienen de la agencia española de noticias EFE y 200.000 del diario El Periódico, y solo 75.000 palabras de Lexesp (Sebastián et al., 2000).

³AnCora-Es es el corpus anotado a diferentes niveles lingüísticos del español más amplio. Se puede descargar gratuitamente en: <http://clic.ub.edu/corpus/ancora>.

⁴Los sufijos *-azo* y *-era* son esencialmente denominales pero los tuvimos en cuenta porque en el trabajo de Santiago and Bustos (1999) aparecen algunos sustantivos deverbales con estos sufijos.

⁵La categorización morfológica sigue el etiquetario Parole (Carmona et al., 1998).

se correspondían con un significado de acción y/o resultado, como por ejemplo ‘cañonazo’ o ‘carrera’. Sin embargo, sí se incluyeron los denominados sustantivos *cousin*, es decir, aquellos sustantivos que si bien no derivan de verbos sí tienen una relación semántica con ellos. Este proceso de selección manual redujo el número de sufijos a 10 (-azo, -era y -dera fueron descartados) y el número de lemas a 817, que son los que finalmente se analizan.

Una vez seleccionada la muestra de análisis, para llevar a cabo el análisis lingüístico de las nominalizaciones deverbales nos centramos en el estudio de las 3.077 ocurrencias correspondientes a los 817 lemas extraídos. En el caso de la estructura argumental se trataba de observar qué constituyentes de los SNs se interpretaban cómo argumentos. Los argumentos que podían ser asociados con las nominalizaciones se consultaban en el léxico AnCora-Verb (Aparicio et al., 2008), asumiendo así la hipótesis de partida de nuestro trabajo: las nominalizaciones deverbales heredan la estructura argumental de sus correspondientes verbos. AnCora-Verb es un léxico que especifica la correspondencia entre las funciones sintácticas, los argumentos y los papeles temáticos de los diferentes verbos teniendo en cuenta la clase semántica de dichos verbos y las alternancias de diátesis en las que participan. Los constituyentes que se podían interpretar como argumentos, se anotaron como tales. A continuación describimos el esquema de anotación utilizado.

3.2. Esquema de anotación

El esquema de anotación seguido es el mismo que fue utilizado en la anotación de la estructura argumental de los verbos en AnCora-Es (Taulé et al., 2008), que a su vez estaba basado en PropBank (Palmer et al., 2005) para la anotación de los argumentos y en VerbNet (Kipper et al., 2000) y (Kipper et al., 2006) para la anotación de papeles temáticos. Usamos el mismo esquema de anotación para sustantivos y verbos porque consideramos que sus argumentos son del mismo tipo, y aún más en el caso de las nominalizaciones deverbales en las que asumimos que heredan la estructura argumental de los verbos. De hecho, nos apoyamos básicamente en el léxico verbal AnCora-Verb para asignar la posición argumental y el papel temático.

El esquema de anotación está formado por un conjunto de 36 etiquetas, la mayoría de las cuales (a excepción de 3) están formadas por una posición argumental y un papel temático. Existen dos etiquetas que solo tienen posición argumental y generalmente se corresponden con los argumentos expresados en el verbo con un complemento preposicional regido. Además, se usa la etiqueta RefMod para aquellos constituyentes que no son argumentos de las nominalizaciones y, por tanto, no pueden recibir etiqueta argumental. Con ella se indica que los constitu-

Etiquetario

yentes que la tienen asignada modifican el nombre al que están complementando pero no constituyen un argumento. Esta etiqueta es exclusiva del etiquetario nominal ya que en los verbos no existen casos de complementos que no constituyan argumentos, como sí los hay en los sustantivos. A continuación, en la Tabla 3.1 mostramos las etiquetas resultantes de la combinación de posición argumental y papeles temáticos (35, en total), a la que cabe añadir la etiqueta RefMod. Téngase en cuenta que cada posición argumental se asocia con unos determinados papeles temáticos.

Argumentos

Al igual que en PropBank los argumentos están numerados de manera incremental —arg0, arg1, arg2, arg3, arg4— expresando el grado de proximidad con el predicado y los argumentos adjuntos, es decir, aquellos que no son exigidos semánticamente por el predicado, se etiquetan como argM. Sin embargo, dado que las etiquetas de PropBank son bastante abstractas y se definen en base a un lexema (se especifican para cada predicado individualmente), nosotros hemos añadido papeles temáticos del tipo de los propuestos en VerbNet con el objetivo de generalizar papeles temáticos en diferentes predicados, siendo estos roles semánticos específicos de una clase o clases de predicados. De hecho, nuestro esquema de anotación es similar a la combinación de las etiquetas semánticas de PropBank y VerbNet propuesta en el proyecto SemLink (Loper et al., 2007; Yi et al., 2007)⁶.

Papeles temáticos

La lista de papeles temáticos que proponemos incluye 19 etiquetas ampliamente utilizadas en lingüística: agt (agente), cau (causa), exp (experimentador), src⁷ (fuente), pat (paciente), tem (tema), atr (atributo), ben (beneficiario), ext (extensión), ins (instrumento), loc (locativo), tmp (tiempo), mnr (manera), ori (origen), des (destino), fin (finalidad), ein (estado inicial), efi (estado final) y adv (adverbial). Usamos estos papeles temáticos porque proporcionan una información semántica más rica de la que proporcionan los argumentos numerados solos. Nuestra propuesta de papeles temáticos se basa en los 23⁸ papeles temáticos de VerbNet ya que estos son suficientemente específicos para nuestros propósitos pero más generales que el gran número de papeles temáticos propuestos en FrameNet (Baker et al., 1998) y (Ruppenhofer et al., 2006). En este recurso los papeles temáticos (elementos del marco, siguiendo su terminología) están organizados

⁶<http://verbs.colorado.edu/semlink/>

⁷Esta abreviatura se corresponde con la palabra inglesa *source*.

⁸De los 23 papeles temáticos utilizados en VerbNet (<http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>), nosotros prescindimos de cuatro: *actor*, *asset*, *stimulus* y *topic*. El primero es un agente inductor propio de construcciones causativas al que nosotros hemos incluido en el papel temático de agente. El papel *asset* ('activo') es específico de una alternancia que en VerbNet es conocida como *sum of money*, 'suma de dinero'; nosotros anotamos las sumas de dinero como extensión. El papel *stimulus* ('estímulo') se encuentra solo en los verbos de percepción; nosotros etiquetamos estos casos con el papel más general de tema. Finalmente, el papel *topic* ('tópico') responde al tema o tópico de conversación de los verbos de comunicación; de nuevo, nosotros anotamos estos casos con el papel más general de tema.

3. ESTRUCTURA ARGUMENTAL DE LAS NOMINALIZACIONES DEVERBALES:
ESTUDIO EMPÍRICO

Argumento	Papel temático	Ejemplo
arg0	agt	La traducción del libro por parte de Juan
	cau	La preocupación de Carlos
	exp	La débil respiración de Laura
	src	Los gritos de María
arg1	tem	La llegada de Andrés
	pat	La construcción de la casa
	loc	El acceso a la ciudad
	∅	La pasión por el fútbol
arg2	loc	La llegada a la meta
	ins	El linchamiento con las porras
	atr	La carencia de talento
	ben	La demostración de fuerza a los allí presentes
	exp	La falta de confianza del equipo
	∅	La fusión con la empresa suiza
	ext	La suma de 20.000 dólares
	efi	Su conversión en la tercera empresa del sector
fin	No se ha encontrado ningún ejemplo con esta etiqueta	
arg3	ori	La salida del país
	ins	El trazo del cadáver con tiza
	atr	El paso del tiempo sin libertad se hace largo
	ben	Un coste elevado para la empresa
	exp	Los antojos de Bárbara en el embarazo
	loc	La alerta de la Dirección en su informe
	ein	La transformación de Luis, de vaqueros a traje
fin	La utilización de la jornada para recoger sugerencias	
arg4	des	El regreso al empleo es complicado
	efi	La transformación de Luis, de vaqueros a traje
argM	adv	La negociación con la oposición
	atr	Un suspiro de alivio
	cau	Críticas por su falta de experiencia
	ext	Ampliación del capital del 16 %
	fin	Apuesta por las patentes para proteger las marcas
	loc	La inversión en investigación en la U.E.
	mnr	La interpretación a su manera de los acuerdos bilaterales
tmp	El triunfo electoral del 10 de junio de 1990	

Tabla 3.1: Conjunto de etiquetas argumentales utilizadas en la anotación de las nominalizaciones deverbales

jerárquicamente y su interpretación es específica para un marco. Sin embargo, los papeles temáticos que nosotros adoptamos son compatibles con los de FrameNet, como muestra el hecho de que en el proyecto SemLink (Palmer, 2009) se han relacionado también los papeles temáticos de FrameNet y VerbNet.

3.3. Estructura argumental: análisis lingüístico

Una vez seleccionada la muestra, un total de 3.077 ocurrencias de nominalizaciones deverbales del subconjunto de 100.000 palabras del corpus AnCora-Es y determinado el esquema de anotación, se procedió a analizar los datos y a la anotación de las mismas. Del resultado de este primer análisis lingüístico se obtuvo la primera versión de la guía de anotación de la estructura argumental de las nominalizaciones deverbales (Sección 8.1.2) e importantes observaciones que están en la base de las reglas heurísticas que nos han permitido anotar automáticamente la estructura argumental de las nominalizaciones del corpus AnCora-Es (Capítulo 4).

El análisis se centraba en todos los constituyentes que formaban parte de los SNs cuyos núcleos eran las 3.077 ocurrencias de la muestra de datos. El objetivo de este análisis era determinar si los constituyentes eran o no argumentales y en el caso que lo fueran, determinar de qué tipo eran y en qué posición se realizan. En este proceso se tenía en cuenta la información sobre la estructura argumental del verbo base correspondiente especificada en el léxico AnCora-Verb. En este sentido, entendemos por argumento de una nominalización aquel constituyente que se pueda interpretar semánticamente como uno de los argumentos asociados al verbo correspondiente. Un argumento es un participante necesario para interpretar el predicado. En cuanto a los complementos no argumentales, nosotros entendemos que son aquellos complementos del nombre que no pueden recibir una interpretación de un participante del predicado, como son por ejemplo los adjetivos calificativos como ‘grande’, ‘pequeño’, ‘precioso’, ‘deplorable’, etc. (1). Aunque en la bibliografía, hay algunos autores que mantienen que los complementos de los sustantivos resultativos no son argumentales (véase el Capítulo 2), nosotros consideramos que todos los tipos de nominalizaciones pueden tener argumentos.

Este tipo de análisis estuvo enfocado a la reflexión y a la obtención de datos sobre la estructura argumental de las nominalizaciones. Este análisis se realizó por dos expertos lingüistas que en todo momento podían comparar las anotaciones y en todos los casos las decisiones eran acordadas. Durante este proceso, hemos obtenido las conclusiones siguientes:

En primer lugar, se observó que no todos los constituyentes que aparecen en los SNs cuyos núcleos son las nominalizaciones pueden siempre expresar sintácticamente argumentos de la nominalización. Los constituyentes de los SNs son: oraciones subordinadas (OSub), SNs, sintagmas adverbiales (SAdv), sintagmas adjetivales (SAs), determinantes posesivos (Poss), pronombres relativos genitivos (GRel) y SPs. Los Poss y los Grel ocupan la posición de especificador del SN mientras que el resto funcionan sintácticamente como complementos del nombre. Entre todos los constituyentes posibles, los que nunca son argumentales son las subordinadas de relativo, ya que siempre especifican una característica del sustan-

3. ESTRUCTURA ARGUMENTAL DE LAS NOMINALIZACIONES DEVERBALES: ESTUDIO EMPÍRICO

tivo pero no expresan un argumento (1).

- (1) Dejar para más tarde el debate sobre [los [grandes]_{SA} **cambios** [*que deben introducirse en el partido*]_{Osub-no argumental}]_{SN}

Los SNs y los SAdvS en la mayoría de las ocasiones no son argumentales pero en algunos casos son constituyentes que pueden expresar algún tipo de argumento adjunto. En los SNs vimos que muchas veces coincide el argumento adjunto de tiempo o lugar con el hecho de que los SNs son entidades con nombre del tipo fecha (3) o lugar (2). Respecto a los SAdvS, se comprobó que en la mayoría de ocasiones no eran argumentales, tal y como propone Meyers (2007). Sin embargo, se observó también que algunos adverbios pueden expresar el mismo tipo de argumento adjunto que expresarían en el caso de los verbos y que suele coincidir con el papel temático adv (adverbial) (4) o mnr (manera) (5).

- (2) [La **concentración** de la producción [en *Europa*_{NE-lugar}]_{SP-argM-loc} cuando los mercados están fuera del continente]_{SN} es un hecho probado.
- (3) [El **anuncio** de la Reina Isabel [en *1985*_{NE-fecha}]_{SP-argM-tmp}]_{SN} sorprendió al mundo.
- (4) [La [*casi*]_{SAdv-argM-adv} **desaparición** de zonas amorfas]_{SN} impide ahora el ataque del oxígeno del aire.
- (5) [La **selección** [*aleatoriamente*]_{SAdv-argM-mnr} de las empresas contratadas]_{SN} ha sido polémica.

En cuanto a los SAs, se observó una importante restricción: solo los adjetivos relacionales pueden interpretarse como argumentos (6) y (7). El resto de SAs son modificadores del nombre y no se les puede asignar argumento alguno (8). Algunos autores (Picallo, 1999) ya habían apuntado este hecho. Téngase en cuenta que los adjetivos relacionales se caracterizan por expresar una relación entre el sustantivo al que complementan y el sustantivo que subyace en su formación derivativa. Por ejemplo, en (6) ‘entramado ideológico’ expresa la relación entre ‘entramado’ e ‘ideas’, y en (7) ‘la innovación empresarial’ se puede parafrasear por ‘la innovación de los empresarios’.

Constituyentes
argumentales

- (6) La precaria situación económica de la organización terrorista y [del **entramado** [*ideológico*]_{SA-arg1-Pat} en el que se sustenta]_{SN} añaden ciertas dosis de credibilidad a las misivas.
- (7) Se está creando un entorno propicio para [la **innovación** [*empresarial*]_{SA-arg0-agt}]_{SN}.
- (8) La visita oficial, en la primera gira del nuevo presidente de Rusia al extranjero, tendrá lugar los próximos días 13 y 14, de acuerdo con [el **comu-**

nicado [*oficial*]_{SA-no argumental}SN.

Los determinantes posesivos, los pronombres de relativo genitivos (cuyo, cuya) y los SPs suelen ser en la mayoría de ocasiones argumentales, si bien existen también SPs no argumentales como complementos de las nominalizaciones (9).

- (9) A través de [un **comunicado** [*de prensa*]_{SP-no argumental}], el presidente señaló que la fusión proporcionará un significativo valor a los accionistas.

Argumentos externos
e incorporados

Además de distinguir entre constituyentes típicamente argumentales y constituyentes no argumentales, otras de las primeras observaciones realizadas es que no siempre los argumentos asociados al verbo base se realizan en el SN de la nominalización correspondiente. En muchas ocasiones los argumentos de las nominalizaciones se encuentran fuera del SN (12), es decir, en el contexto oracional o textual de la nominalización. En otras ocasiones, el argumento está incorporado en la misma nominalización (11). Aunque nuestro trabajo se centra en los argumentos dentro del SN (10), los argumentos incorporados se anotan, al ser pocos, en el proceso de validación manual descrito en el Capítulo 8 de este trabajo, y los argumentos externos al SN, aunque su tratamiento es incipiente, los abordaremos en el Capítulo 10 como una línea futura de trabajo.

- (10) [La **construcción** [*de la casa*]_{SP-arg1-pat} [*por parte de Juan*]_{SP-arg0-agt}]_{SN} duró dos años.
- (11) [El **invento**_{arg1-pat} [*de Juan*]_{SP-arg0-agt}]_{SN} tuvo mucho éxito.
- (12) [*Juan*]_{arg0-agt} tomó [la **decisión** más acertada]_{SN}.

En el ejemplo (10) los dos argumentos, el argumento paciente (arg1-pat) y el argumento agente (arg0-agt), se realizan por SPs dentro del SN: ‘de la casa’ (paciente), ‘por parte de Juan’ (agente). En el ejemplo (11), ‘invento’ tiene el argumento paciente (arg1-pat) incorporado en el mismo nombre, mientras que el argumento agente (arg0-agt) se realiza por un SP ‘de Juan’. Este SN se puede parafrasear por la oración ‘Juan inventó un invento’. En el ejemplo (12), ‘Juan’ es semánticamente el argumento agente (arg0-agt) de ‘decisión’, pero se vincula al sustantivo mediante el verbo soporte tomar, y por lo tanto, se encuentra fuera del SN.

En cuanto a los constituyentes argumentales, se obtuvieron las siguientes observaciones :

Argumentos internos
Posesivos

Los determinantes posesivos que especifican las nominalizaciones deverbales suelen expresar algún tipo de argumento de la nominalización y muestran una preferencia bastante clara por realizar el argumento equivalente al sujeto del verbo del que deriva dicha nominalización, por lo que el tipo de argumento asociado (arg1, arg0) varía en función de la clase semántica asociada al verbo base de la nominalización (13), (14).

3. ESTRUCTURA ARGUMENTAL DE LAS NOMINALIZACIONES DEVERBALES:
ESTUDIO EMPÍRICO

(13) [[*Su*]_{Poss-arg0-agt} **disposición** constante a hacer el bien]_{SN}.

(14) [[*Su*]_{Poss-arg1-tem} **entrada** en la sala]_{SN} tranquilizó al presidente.

Los pronombres de relativo genitivos (*cuyo*, *cuya*) también pueden expresar argumentos de las nominalizaciones, sin embargo, dado que el número de ejemplos con argumentos realizados mediante este constituyente es escaso no podemos apuntar ninguna preferencia clara por un tipo determinado de argumento, a lo sumo cabría destacar que solo *arg1* y *arg2* han sido realizados por este constituyente en la muestra analizada (15).

Relativos genitivos

(15) Hemos de reconocer un don o talento natural [[*cuya*]_{GRel-arg1-tem} **carencia**]_{SN} ninguna educación puede suplir.

El constituyente SP es el más frecuente como argumento de las nominalizaciones deverbales, aunque no todos los SPs son argumentos de las nominalizaciones, como ocurría en (9). El tipo de argumento asociado con los SPs depende, en gran medida, de la preposición que introduce el SP. Existen algunas preposiciones que tienen un valor semántico concreto y, por lo tanto, se asocian con argumentos específicos. Por ejemplo, la preposición ‘*hacia*’, normalmente introduce un SP que marca un destino (*arg4-des*) (16), mientras que ‘*desde*’ indica un origen (*arg3-ori*) (17). De la misma manera, la preposición ‘*para*’ introduce normalmente un SP que expresa finalidad (*argM-fin*) (18), ‘*durante*’ o ‘*tras*’ normalmente expresan tiempo (*argM-tmp*) (19), (20) y ‘*según*’ o ‘*sin*’ implican un argumento adverbial (*argM-adv*) (22),(21).

SPs

(16) En [la **marcha** [*hacia* Bruselas]_{SP-arg4-des}]_{SN} fue cortando cabezas.

(17) Ha sido muy importante la recuperación de Hierro, un hombre vital en [la **salida** [*desde* atrás]_{SP-arg3-ori}][*con* el balón]_{SP-argM-mnr}]_{SN}.

(18) Es preciso aplicar [**remedios** serios [*para* restablecer la competencia]_{SP-argM-fin}]_{SN}.

(19) [El **incremento** [*del* número de desempleados]_{SP-arg1-tem} [*durante* el pasado mes de Mayo]_{SP-argM-tmp}]_{SN} se debe al aumento de la población activa.

(20) [La **caída** [*del* gobierno]_{SP-arg1-tem} [*tras* las manifestaciones]_{SP-argM-tmp}]_{SN} ha sido bien recibida por la comunidad internacional.

(21) [La **matanza** [*sin* escrúpulos]_{SP-argM-adv} [*de* niños]_{SP-arg1-tem}]_{SN} conmocionó al país.

(22) [La **compra** [*según* las normas]_{SP-argM-adv}]_{SN} resultó insuficiente.

Otra característica que observamos en los SPs es que si el SP de la nominaliza-

ción se introduce con la misma preposición que un complemento del verbo base del que deriva o con el que se relaciona semánticamente, entonces el argumento asociado a la nominalización coincide con el argumento asignado al complemento verbal con la mencionada preposición. Por ejemplo, el verbo ‘combatir’ en (23) tiene un argumento instrumento (arg2-ins) expresado por un SP introducido por la preposición ‘con’. Si el sustantivo ‘combate’ aparece con un SP también introducido por la misma preposición, el tipo de argumento asociado con ese SP suele ser el mismo (24).

(23) Necesitaban **combatirlo** [*con las armas*]SP-arg2-ins.

(24) [El **combate** [*con la espada*]SP-arg2-ins]SN siempre es más elegante.

Sin embargo, los SPs no marcados (los introducidos por *de* y, en general, aquellos en que la preposición no aporta ningún significado específico o no existe una relación con un complemento verbal preposicional) no mostraron ninguna preferencia clara por ningún tipo de argumento en concreto: encontramos tanto SPs que expresan argumentos principales (arg1, arg0, arg2, arg3, arg4) como argumentos adjuntos (argM). Por ejemplo, en (25) el SP introducido por *de* expresa un argumento con el papel temático de paciente (arg1-pat) y en el ejemplo (26) otro SP introducido por la misma preposición expresa un argumento adjunto con el papel temático de extensión (argM-ext). Los papeles temáticos se establecen en función de los que tiene asociado el verbo base a esa misma posición argumental. A pesar de todo, sí que observamos que, entre los argumentos principales, el arg1 era el que se realizaba con más frecuencia; esto se explica porque el arg1 corresponde generalmente con los papeles de paciente y tema, es decir, los argumentos que se corresponden con los objetos directos de verbos transitivos y sujetos de verbos inacusativos, y son los argumentos más decisivos para la comprensión del predicado. A este argumento le siguen el arg0 y el arg2, que son los otros argumentos más próximos al predicado, mientras que el arg3 y el arg4 eran casi residuales.

(25) Dinámica de deterioro y [**deslegitimación** [*de las instituciones*]SP-arg1-pat]SN.

(26) Los beneficios económicos inmediatos supondrán al menos [un **ahorro** [*de 150 millones de dólares*]SP-argM-ext]SN.

Los SA relacionales, es decir, aquellos SAs que tienen como núcleo un adjetivo relacional y que son los únicos que pueden ser interpretados como argumentales, como hemos visto, expresan con más frecuencia el arg0, a diferencia de los SPs. Sin embargo, esto se debe a que la mayor parte de las veces aparecen con un SP que expresa el arg1 (27). Cuando el único argumento es un SA relacional el argumento tiende a ser arg1 (6) o arg0 (7) con la misma frecuencia.

- (27) [La **decisión** [empresarial]_{SA-arg0-agt}] [de vender los activos]_{SP-arg1-pat}]SN no ha sentado bien a los accionistas.

Finalmente, observamos que si aparecen dos constituyentes del mismo tipo (SPs ^{SAs + SPs} o SAs) no existe una tendencia clara respecto a qué argumentos y en qué orden se realizan, aunque la mayoría de las veces uno de los dos argumentos expresados es el arg1 (si la clase verbal de la que deriva la nominalización no es inergativa) y el otro puede ser tanto un arg0 (28) como un arg2 (29) (en función de la clase verbal origen) como un argumento adjunto (30). Evidentemente, si la clase verbal de la que deriva la nominalización es inergativa los argumentos realizados son el arg0 y un argumento adjunto, aunque hay muy pocos casos con esta configuración en la muestra analizada (30).

- (28) Fue [un **lanzamiento** [de falta]_{SP-arg1-pat}] [de Alonso]_{SP-arg0-agt}]SN.
(29) Los inversores comenzarán a tomar posiciones por [la **entrada** [de Terra]_{SP-arg1-tem}] [en el Ibex-35]_{SP-arg2-loc}] [el próximo día 31]_{SP-argM-tmp}]SN.
(30) [El **regreso** [del Real Madrid]_{SP-arg0-agt}] [el jueves]_{SP-argM-tmp}]SN.

3.4. Conclusiones

En este capítulo se ha detallado cómo a partir de la observación de la casuística presentada en la Sección 3.1, se han obtenido una serie de observaciones lingüísticas sobre la estructura argumental de las nominalizaciones deverbales del español (Sección 3.3). A partir de estas observaciones hemos generalizado y hemos elaborado un conjunto de Reglas Heurísticas para las Nominalizaciones deverbales (RHN) que nos han permitido anotar automáticamente la estructura argumental de las nominalizaciones en la totalidad del corpus AnCora-Es. Este sistema automático se presenta en el capítulo siguiente. En este capítulo también se ha presentado el esquema de anotación seguido (Sección 3.2), que será asimismo utilizado para la anotación automática.

CAPÍTULO 4

ANOTACIÓN AUTOMÁTICA DE LOS ARGUMENTOS INTERNOS

En este capítulo presentamos la metodología seguida para anotar automáticamente los argumentos internos de las nominalizaciones deverbales en el corpus AnCora-Es. Describimos y evaluamos el conjunto de reglas heurísticas, que hemos llamado RHN, y las hipótesis lingüísticas que las sustentan, desarrolladas para asignar la posición argumental y el papel temático a los argumentos de las nominalizaciones deverbales. Primero mostramos cómo se ha llevado a cabo la anotación automática de las nominalizaciones deverbales, las reglas heurísticas implementadas y los recursos que utilizan (Sección 4.1). A continuación, presentamos la evaluación de las reglas heurísticas a partir de la validación manual (Véase el Capítulo 8) llevada a cabo (Sección 4.2). Finalmente, en la Sección 4.4 presentamos las conclusiones.

4.1. Reglas Heurísticas y Recursos Lingüísticos

El proceso de anotación de la estructura argumental de las nominalizaciones deverbales en el corpus AnCora-Es (Peris and Taulé, 2011b) se llevó a cabo en dos etapas (Figura 4.1): la primera consiste en la anotación automática que describimos en este capítulo y la segunda en la validación manual descrita en el Capítulo 8, que permite la evaluación de la anotación automática.

Las nominalizaciones deverbales candidatas a ser anotadas se obtuvieron de manera semiautomática, utilizando el mismo proceso de extracción que en el estudio empírico pero aplicado a la totalidad del corpus AnCora-Es (500.000 palabras). A partir del conjunto predefinido de sufijos nominalizadores (*-a*, *-aje*, *-ión/-*

Muestra analizada

ción/-sión/-ón, -da/-do, -dura/-ura, -e, -ido, -miento/-mento, -ncia/-nza, -o/-eo) que toman como bases de la derivación verbos y tienen un significado de acción-resultado (Santiago and Bustos, 1999), se seleccionaron manualmente una lista de nominalizaciones con un significado potencialmente deverbal en la que también se especifica el verbo base de cada una de las nominalizaciones. En total, se seleccionaron 1.655 nominalizaciones que se corresponden con un total de 24.864 ocurrencias en el corpus AnCora-Es.

RHN

La anotación automática de las 1.655 nominalizaciones de la lista se llevó a cabo a partir de un conjunto de reglas heurísticas creadas manualmente que codifican gran parte del conocimiento lingüístico obtenido en el estudio empírico descrito en el capítulo anterior. Este paquete de Reglas Heurísticas para las Nominalizaciones deverbales se ha llamado RHN y es con este acrónimo que nos referiremos a ellas a partir de ahora. El objetivo principal de RHN es proyectar la estructura argumental de los verbos declarada en el léxico verbal AnCora-Verb sobre las nominalizaciones deverbales que les corresponden. Para lograr este objetivo, RHN, además de utilizar información de dicho léxico verbal, también usa información obtenida del corpus AnCora-Es y de una lista de adjetivos relacionados creada manualmente.

Para la anotación de los argumentos internos al SN, hemos seguido el esquema de anotación especificado en el capítulo anterior. Recuérdese que es el mismo esquema que fue utilizado en la anotación de la estructura argumental de los verbos en AnCora-Es, que a su vez estaba basado en PropBank y en VerbNet. De esta manera, además, aseguramos la consistencia en la anotación de argumentos de varios predicados (verbos y nominalizaciones), haciendo que los recursos desarrollados con este esquema de anotación sean compatibles con los del inglés. Usamos el mismo esquema de anotación para sustantivos y verbos porque creemos que sus argumentos son del mismo tipo y, especialmente, en el caso de las nominalizaciones deverbales en las que nos apoyamos básicamente en el léxico verbal AnCora-Verb para asignar la posición argumental y el papel temático. Las reglas heurísticas desarrolladas solo utilizan 26 de las 36 etiquetas semánticas posibles para anotar automáticamente los argumentos y papeles temáticos de las nominalizaciones. Las diez restantes son poco frecuentes y no se consideró eficiente desarrollar reglas heurísticas para su anotación. Los constituyentes que no pueden ser argumentos son los SA no relacionales, los SNs, los SAdv y las Oraciones Subordinadas, que reciben la etiqueta RefMod, como veremos, mientras que sí pueden ser argumentales los SPs, los SA relacionales, los GRel y los Poss, como se vio en el estudio empírico.

A continuación describimos los recursos lingüísticos utilizados y las reglas heurísticas desarrolladas a partir de estos recursos.

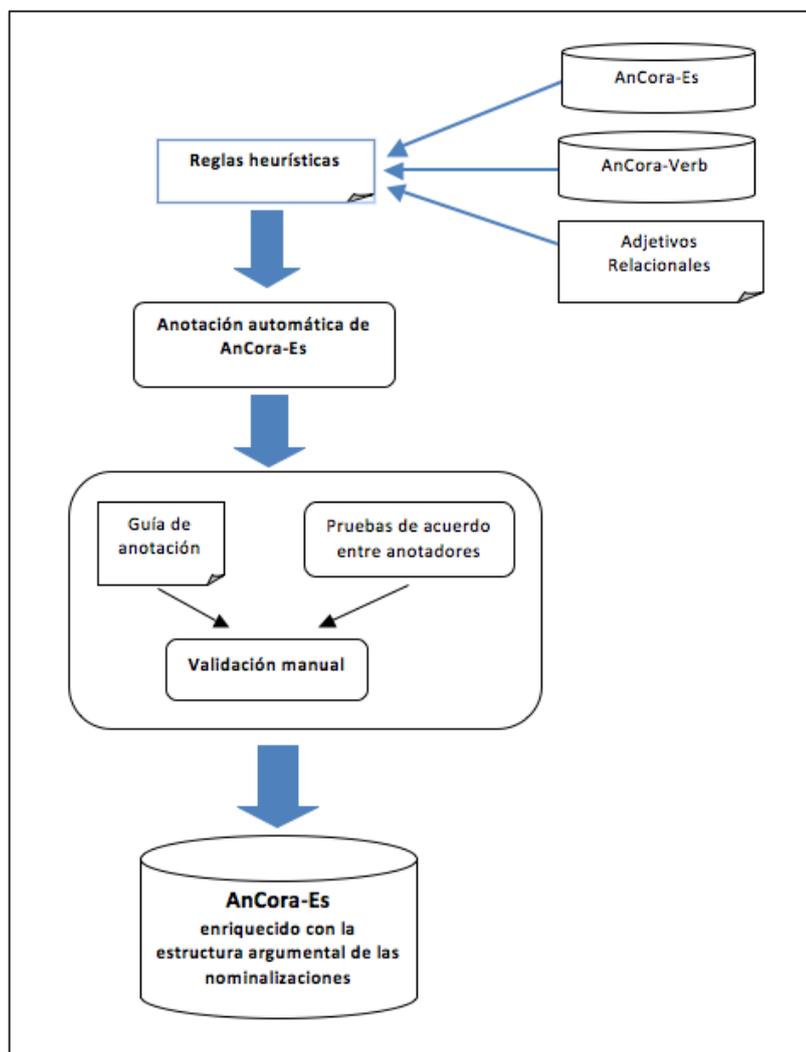


Figura 4.1: Proceso de anotación de la estructura argumental de las nominalizaciones deverbales en AnCora-Es

4.1.1. Recursos Lingüísticos

En esta subsección presentamos los recursos lingüísticos utilizados en el proceso automático: el corpus AnCora-Es, el léxico verbal AnCora-Verb y la lista de adjetivos relacionales.

AnCora-Es es un corpus del español de 500.000 palabras cuyas características se han especificado en el capítulo anterior. Cabe resaltar que este corpus juega dos papeles en este proceso de anotación automática: por una parte, es el corpus que anotamos, y por otra, es uno de los recursos lingüísticos que usamos para

llevar a cabo la anotación automática; por ejemplo, los tipos de constituyentes y las etiquetas de las entidades nombradas son informaciones que usamos en las reglas heurísticas (Véase 4.1.2).

AnCora-Verb-Es

AnCora-Verb-Es es un léxico que contiene 2.830 lemas de verbos del español que se corresponden con aquellas apariciones verbales del corpus AnCora-Es. En este léxico se especifica la correspondencia entre las funciones sintácticas, los argumentos y los papeles temáticos de los diferentes verbos teniendo en cuenta la clase semántica de dichos verbos y las alternancias de diátesis en las que participan. Un verbo puede tener diferentes sentidos y cada uno de ellos puede estar relacionado con una o más clases semánticas. Estas clases semánticas pertenecen a cuatro grupos que se definen teniendo en cuenta las cuatro clases eventivas propuestas por Vendler (1967) y Dowty (1979) –realizaciones, logros, estados y actividades– y las alternancias de diátesis (Vázquez et al., 2000). Las clases semánticas principales se subdividen en doce subclases más específicas, convenientemente resumidas en la Tabla 4.1.

Clase A: Realizaciones	Clase B: Logros
A1: Causativa-Transitiva: ‘dañar’	B1: Cambio de lugar: ‘llegar’
A2: Agentiva-Transitiva : ‘acatar’	B2: Cambio de estado: ‘convertir’
A3: Agentiva-Ditransitiva : ‘enviar’	
Clase C: Estados	Class D: Actividades
C1: Estado existencial: ‘marchitar’	D1: Agentiva-inergativa: ‘trabajar’
C2: Estado atributivo: ‘ser’	D2: Experimental-inergativa: ‘vivir’
C3: Estado escalar: ‘costar’	D3: Fuente-inergativa: ‘llorar’
C4: Estado beneficiario: ‘gustar’	

Tabla 4.1: Clases semánticas verbales

De este recurso lingüístico tenemos en consideración dos tipos de información para desarrollar las reglas heurísticas (Subsección 4.1.2): (i) la preposición que es núcleo de los SPs que son complementos verbales, puesto que su etiqueta argumental puede ser proyectada sobre los SPs complementos de las correspondientes nominalizaciones si comparten la preposición; y (ii), la clase semántica verbal, que proporciona la base lógica para la asignación de posición argumental y papeles temáticos a los argumentos de las nominalizaciones deverbales.

Adjetivos Relacionales

Finalmente, dado que los adjetivos relacionales son los únicos que pueden ser interpretados como argumentos de las nominalizaciones deverbales (Picallo, 1999; Bosque and Picallo, 1996), creamos automáticamente un lista de adjetivos relacionales potenciales extrayendo de AnCora-Es los adjetivos que terminaban en *-al*, *-ario*, *-es*, *-ico*, *-ista*, *-stico* (Rainer, 1999). Los adjetivos relacionales se

caracterizan por su posición detrás de la nominalización y por expresar una relación entre la nominalización (‘actuación’) y un sustantivo a partir del cual se deriva el adjetivo relacional (‘policía’ > ‘policial’ en ‘actuación policial’). Tras la obtención de esta lista de adjetivos relacionales potenciales, seleccionamos manualmente los adjetivos que realmente eran relacionales (331) de los 746 lemas adjetivales automáticamente obtenidos¹. En el Apéndice A se puede encontrar la lista definitiva de adjetivos relacionales.

4.1.2. Reglas Heurísticas

Para anotar la estructura argumental de las nominalizaciones deverbales en Ancora-Es, construimos manualmente un paquete de 107 reglas heurísticas (RHN) cuyo objetivo es el de asociar los constituyentes de los SNs de núcleo de verbal con su correspondiente posición argumental y papel temático usando los recursos lingüísticos mencionados. RHN incorpora el conocimiento lingüístico obtenido a partir del estudio empírico sobre la estructura argumental de las nominalizaciones deverbales, por lo que su evaluación, supone también la evaluación de las hipótesis lingüísticas subyacentes. Las reglas se organizan en una estructura de lista de decisión, es decir, se intentan aplicar secuencialmente hasta que una de ellas se aplica con éxito. El objetivo de aplicación de las reglas son los SNs constituidos por una nominalización (N) y un CONTEXTO particular, que puede comprender uno, dos, tres o más constituyentes. Las reglas son del tipo “ *si* <condición> *entonces* <acción>”, donde la <condición> es una combinación lógica de predicados sobre N y su contexto (denotado por la variable X) y la <acción> es la etiqueta semántica que se le asigna (posición argumental y papel temático). Un ejemplo de la sintaxis de las reglas se muestra a continuación:

```
“dentro_de (X, CONTEXTO) Y
(tipo_de (X,SN) O tipo_de (X,SP)) Y
entidad con nombre (X) Y
tipo_de_entidad con nombre (X,Lugar) >argM-loc ”
```

La regla anterior se lee de la siguiente manera: si X es un complemento dentro del contexto (SN) de una nominalización, y este complemento es del tipo SN o SP y además este complemento se corresponde con una entidad con nombre de lugar (X,Lugar), a ese complemento se le asocia el argumento adjunto de lugar (argM-loc).

¹La idea inicial para detectar los adjetivos relacionales era adaptar al español el clasificador de adjetivos desarrollado por Boleda (2007), pero el coste en esfuerzo y tiempo de este proyecto para anotar la estructura argumental de las nominalizaciones deverbales no valía la pena.

En RHN se distinguen dos tipos de reglas: (i) 14 reglas generales basadas en la información lingüística codificada en AnCora-Es, y (ii) 93 reglas específicas que tienen en cuenta, además, la información declarada en el léxico verbal AnCora-Verb.

RHN: Reglas Generales.

Estas reglas se aplican en primer lugar y están basadas en la información semántica, morfosintáctica y léxica anotada en AnCora-Es. Estas reglas permiten asignar inequívocamente una posición argumental y papel temático a un constituyente de un SN de núcleo deverbal. Diferenciamos tres tipos de reglas generales en función de la información que tienen en cuenta (Véase la Tabla 4.2).

Reglas de Entidad con Nombre	NE _[Lugar] > {SN/SP}-argM-loc NE _[Fecha] > {SN/SP}-argM-tmp
Reglas de Preposición	SP _[durante] > SP-argM-tmp SP _[tras] > SP-argM-tmp SP _[para] > SP-argM-fin SP _[sin] > SP-argM-adv SP _[según] > SP-argM-adv SP _[hacia] > SP-arg4-des SP _[desde] > SP-arg3-ori SP _[mediante] > SP-argM-mnr
Reglas de Constituyente	S > RefMod SAdv > RefMod SA _[no-relacional] > RefMod SN _[no-entidad.con.nombre] > RefMod

Tabla 4.2: Notación simplificada de las Reglas Generales

Reglas de Entidad con Nombre

a) Reglas de Entidad con Nombre: el primer tipo de regla general tiene en cuenta la información semántica que contienen las entidades con nombre, *Named Entities (NE)* de “lugar” o “fecha”. Asumimos que los SNs y SPs que las contienen se corresponden con argumentos adjuntos de lugar y tiempo: argM-loc (1) y argM-tmp (2), respectivamente.

- (1) Agilizar los trámites para responder a [la **falta** de mano de obra [en *Cataluña*_{NE-lugar}]_{SP-argM-loc}]_{SN}.
- (2) La compañía presentó una auditoría limpia por primera vez desde [su **constitución** [en *1989*_{NE-fecha}]_{SP-argM-tmp}]_{SN}.
- (3) Presentaron el acto con momentos emblemáticos y con [**anuncios** [(La

Lechera, Telefunken)]SN-RefMod]SN.

Por lo tanto, solo los SNs que constituyen una entidad con nombre de “lugar” o “fecha” son anotadas como argumentos. El resto de SNs no se consideran argumentos de una nominalización deverbal (Meyers, 2007) y por eso no reciben una etiqueta argumental. En estos casos, los SNs complementos de nominalizaciones se anotan con la etiqueta RefMod, que indica que modifican la referencia de la nominalización (3).

b) Reglas de preposición: el segundo tipo de regla general tiene en cuenta información léxica, concretamente el tipo de preposición que encabeza los SPs ya que algunas pueden ser indicadoras de un papel temático específico, tal y como vimos en la Sección 3.3. Por ejemplo, la preposición ‘hacia’, normalmente introduce un SP que denota un destino (4), mientras que ‘desde’ puede indicar un origen (5). De la misma manera, la preposición ‘para’ normalmente introduce una finalidad (6) y ‘durante’ un argumento temporal (7).

- (4) Su posición en la general le permitió [una **marcha** triunfal [*hacia* la meta]SP-arg4-des]SN.
- (5) La supresión de [**vuelos** [*desde* Barcelona]PP-arg3-ori][a Atlanta y Nueva York]SP-arg4-des]SN no es una consecuencia directa de los ataques terroristas.
- (6) Vio difícil [la **negociación** [*para* la renovación del Concierto Económico]SP-argM-fin]SN.
- (7) Ha sido condenado a [la **prohibición** [de la licencia de circulación]SP-arg1-pat [*durante* un año]SP-argM-tmp]SN.

Sin embargo, no siempre estas hipótesis resultan ciertas. Los SPs introducidos por ‘desde’, por ejemplo, no siempre indican origen (5) sino que muchas veces también denotan argumentos temporales (8) (Véase la Sección 4.2).

- (8) Ha crecido el gasto en los hogares, [el primer **incremento** [*desde* hace siete meses]SP-argM-tmp]SN.

c) Reglas de constituyente: el tercer y último tipo de reglas generales tiene en cuenta información morfosintáctica, en concreto, el tipo de constituyente que modifica las nominalizaciones deverbales: las Osub, los SAdv, los SNs que no contienen entidades con nombre y los SAs que no tienen como núcleo un adjetivo relacional. Respecto a las oraciones subordinadas y la mayoría de SAdv (Badia, 2002) y (Meyers, 2007), se considera que no son argumentales, por lo que se les asigna la etiqueta RefMod (9), (10). A pesar de esto, observamos en el estudio empírico (Capítulo 3) que algunos SAdv pueden ser también argumentos

adjuntos de las nominalizaciones deverbales (11), pero dado que no había manera automática de distinguirlos de los no argumentales, optamos por asignar por defecto la etiqueta RefMod a todos los SAdv.

- (9) Podía estar tras [las **amenazas** [*que he recibido*]_{Osub-RefMod}]_{SN}.
- (10) Quieren [una **investigación** [*complementaria*]_{SA-RefMod} [dentro del sumario sobre la muerte de Diana de Gales]_{SAdv-RefMod}]_{SN}.
- (11) Protagonizó [un **recorrido** [*a pie*]_{SAdv-argM-mnr} [por la Rambla]_{SP-argM-loc}]_{SN}.

En cuanto a los SAs, es comúnmente aceptado que solo los adjetivos relacionales (12) pueden ser interpretados como argumentos de las nominalizaciones deverbales (Picallo, 1999; Grimshaw, 1990; Bosque and Picallo, 1996). Los adjetivos relacionales se diferencian de los atributivos en que solo estos últimos expresan una cualidad del nombre y pueden aparecer tanto delante (13) como detrás del nombre (10). Por lo tanto, solo los adjetivos relacionales de la lista creada se anotan como argumentos de las nominalizaciones deverbales, los restantes se etiquetan como RefMod.

- (12) El tema de conversación era [la **actuación** [*policial*]_{SA-arg0-agi}]_{SN}.
- (13) Hoy, tras [una [*maratoniana*]_{SA-RefMod} **negociación** [de trece horas]_{SP-argM-tmp}]_{SN}, se ha aprobado un nuevo texto sobre la reforma del seguro de desempleo.

RHN: Reglas Específicas.

Estas reglas se diseñaron para ser aplicadas tras las reglas generales, por lo que no tienen en cuenta los constituyentes que ya se han asignado mediante las reglas generales. Se basan en la información especificada en el léxico AnCora-Verb, del cual se obtiene la clase semántica verbal y la preposición que introduce los complementos verbales preposicionales. La clase semántica verbal nos permite asignar argumentos y papel temático a los constituyentes de los SNs de núcleo deverbal, mientras que la preposición permite proyectar el argumento y papel temático de los SPs argumentales de los verbos sobre los SPs de los SNs de núcleo deverbal con los que comparten preposición. Cabe recordar aquí que consideramos un total de 12 clases semánticas que se organizan alrededor de los cuatro tipos eventivos—realizaciones, logros, estados y actividades (Vendler, 1967; Dowty, 1979): las clases A se corresponden con las realizaciones, las clases B con los logros, las clases C con los estados y las clases D con las actividades (Véase la Tabla 4.1).

Es importante indicar también que la correspondencia entre los argumentos de los verbos y los de las nominalizaciones deverbales se garantiza por la lista

de nominalizaciones deverbales candidatas a ser anotadas en la que se establece para cada una de ellas el verbo base que le corresponde. Sin embargo, fue necesario considerar si el verbo correspondiente tenía uno o más significados. Si el verbo es monosémico (solo se le asocia un sentido y, por tanto, una única clase semántica), entonces las reglas tienen en cuenta la información de ese sentido. Si el verbo es polisémico, entonces el sentido verbal que se corresponda con la clase semántica con el mayor número de argumentos es elegido automáticamente y las reglas toman la información de este sentido. De esta manera, un mayor número de argumentos están disponibles para ser proyectados.

Las reglas específicas también tienen en cuenta el número y tipo de constituyentes de los SNs de núcleo de verbal (SP, SA, GRel, Poss). Dependiendo de cuántos constituyentes tiene el SN de núcleo de verbal, los argumentos verbales proyectados varían. La información sobre el tipo de constituyente también es importante puesto que algunos argumentos verbales prefieren proyectarse en un tipo de constituyente específico. Por ejemplo, los determinantes posesivos parecen preferir interpretarse como los argumentos correspondientes a los sujetos verbales. Consideramos dos tipos de reglas específicas: a) reglas de un único constituyente, y b) reglas de dos o más constituyentes. Las primeras se resumen en la Tabla 4.3 y las segundas en las Tablas 4.5, 4.6, 4.7, 4.8 y 4.9.

Describimos a continuación las reglas de un solo constituyente que, recuerde, solo afectan a aquellos constituyentes que pueden ser argumentales (SPs, SAs, GRel, Poss). Reglas de un constituyente

a1) Las reglas que tienen en cuenta los SPs se basan en dos supuestos. En primer lugar, tenemos como hipótesis que un SP que modifica a una nominalización tiene el mismo argumento y papel temático que un SP complemento del verbo base correspondiente, si comparten la preposición. Por ejemplo, ‘experimentar’ tiene como complemento un SP arg2 instrumento (SP-arg2-ins) generalmente introducido por la preposición ‘con’ (14); por lo tanto, en la nominalización de verbal correspondiente, ‘experimento’, se asigna el mismo argumento y papel temático al SP introducido por la misma preposición ‘con’ (15). Reglas de SP

- (14) Denis Papin se dedicó a **experimentar** [*con* el vapor de agua y la marmita que lleva su nombre]_{SP-arg2-ins}.
- (15) Las tropas japonesas llevaron a cabo [**experimentos** [*con* armas bacteriológicas]_{SP-arg2-ins}]_{SN}.

En segundo lugar, observamos que los SPs introducidos por la preposición ‘de’, la preposición no marcada del español, mostraban una tímida preferencia por la interpretación de arg1 (16) siempre y cuando este argumento esté presente en la estructura eventiva del verbo correspondiente, esto es, en las clases semánticas verbales A, B y C, pero no D. En el caso de las nominalizaciones derivadas de

Reglas de SP	<p>N+SP_[prepn] y V+SP_[prepn] arg-th-rolen >N+SP-arg-th-rolen</p> <p>N+SP_[de] si la clase semántica verbal es A1 >SP-arg1-tem</p> <p>N+SP_[de] si la clase semántica verbal es A2 >SP-arg1-pat</p> <p>N+SP_[de] si la clase semántica verbal es A3 >SP-arg1-pat</p> <p>N+SP_[de] si la clase semántica verbal es B >SP-arg1-tem</p> <p>N+SP_[de] si la clase semántica verbal es C >SP-arg1-tem</p> <p>N+SP_[de] si la clase semántica verbal es D1 >SP-arg0-agt</p> <p>N+SP_[de] si la clase semántica verbal es D2 >SP-arg0-exp</p> <p>N+SP_[de] si la clase semántica verbal es D3 >SP-arg0-src</p>
Reglas de SA	<p>N+SA si la clase semántica verbal es A1 >SA-arg1-tem</p> <p>N+SA si la clase semántica verbal es A2 >SA-arg1-pat</p> <p>N+SA si la clase semántica verbal es A3 >SA-arg1-pat</p> <p>N+SA si la clase semántica verbal es B >SA-arg1-pat</p> <p>N+SA si la clase semántica verbal es C >SA-arg1-pat</p> <p>N+SA si la clase semántica verbal es D1 >SA-arg0-agt</p> <p>N+SA si la clase semántica verbal es D2 >SA-arg0-exp</p> <p>N+SA si la clase semántica verbal es D3 >SA-arg0-src</p>
Reglas de GRel	<p>GRel+N si la clase semántica verbal es A1 >GRel-arg1-tem</p> <p>GRel+N si la clase semántica verbal es A2 >GRel-arg1-pat</p> <p>GRel+N si la clase semántica verbal es A3 >GRel-arg1-pat</p> <p>GRel+N si la clase semántica verbal es B >GRel-arg1-pat</p> <p>GRel+N si la clase semántica verbal es C >GRel-arg1-pat</p> <p>GRel+N si la clase semántica verbal es D1 >GRel-arg0-agt</p> <p>GRel+N si la clase semántica verbal es D2 >GRel-arg0-exp</p> <p>GRel+N si la clase semántica verbal es D3 >GRel-arg0-src</p>
Reglas de Poss	<p>Poss+N si la clase semántica verbal es A1 >Poss-arg0-cau</p> <p>Poss+N si la clase semántica verbal es A2 >Poss-arg0-agt</p> <p>Poss+N si la clase semántica verbal es A3 >Poss-arg0-agt</p> <p>Poss+N si la clase semántica verbal es B >Poss-arg1-tem</p> <p>Poss+N si la clase semántica verbal es C >Poss-arg1-tem</p> <p>Poss+N si la clase semántica verbal es D1 >Poss-arg0-agt</p> <p>Poss+N si la clase semántica verbal es D2 >Poss-arg0-exp</p> <p>Poss+N si la clase semántica verbal es D3 >Poss-arg0-src</p>

Tabla 4.3: Notación simplificada de las reglas específicas de un constituyente

verbos de la clase D, los SPs solo pueden ser interpretados como arg0 ya que es el único argumento posible de esta clase de verbos (17).

- (16) Pujol dio un toque de alerta sobre [el **aumento** [*de* los accidentes laborales]_{SP-arg1-tem}]_{SN}.

- (17) La gran novedad en la lista es [el **regreso** [*de* Richard Dutruel]_{SP-arg0-agt}]_{SN}.

Por lo tanto, en el caso de los SPs, las reglas consideran primero la preposición y luego la clase semántica del verbo. Por ejemplo, si la preposición es ‘de’, el argumento y el papel temático será arg1-tem si el verbo pertenece a las clases A1, B o C, arg1-pat si el verbo pertenece a las clases A2 o A3 y arg0-agt, arg0-exp y arg0-src si el verbo pertenece a las clases D1, D2, o D3, respectivamente. En el ejemplo (16) el verbo base de la nominalización ‘aumento’, ‘aumentar’, pertenece a la clase semántica B1 por lo que el argumento y papel temático asociado al SP es arg1-tem. En cambio, en el ejemplo (17) el verbo base de ‘regreso’, ‘regresar’, es de la clase semántica D, siendo arg0-agt el argumento y papel temático asociado al SP. Si la preposición no es ‘de’, las reglas buscan en la entrada del verbo base del léxico AnCora-Verb un argumento introducido por la misma preposición. Si se encuentra, el argumento y papel temático asociado a dicho complemento verbal se asigna también al complemento nominal. Si no se encuentra ningún SP con la misma preposición, se asigna la etiqueta por defecto argM.

a2) Las reglas que tratan los SAs (19) y los GRels (18) cuando aparecen en solitario en el SN, dado que son constituyentes que no mostraron una preferencia clara por ninguna configuración, siguen las mismas reglas que los SPs introducidos por ‘de’ para la asignación de argumento y papel temático, es decir, se interpretan como arg1 cuando el verbo base pertenece a las clases semánticas verbales A, B y C (18), y como arg0 si la nominalización se deriva de un verbo de la clase D (19). Cabe recordar que los SAs a los que nos referimos en estas reglas tienen como núcleos adjetivos pertenecientes a la lista de adjetivos relacionales y aparecen tras la nominalización, es decir, son potencialmente argumentales.

Reglas de SA
Reglas de Grel

- (18) Más de 1.200 candidatos se presentarán a las elecciones [[*cuya*]-arg1-pat **celebración**]_{SN} será en mayo.

- (19) Se está creando un entorno propicio para [la **innovación** [*empresarial*]_{SA-arg0-agt}]_{SN}.

En el ejemplo (18) tenemos que el verbo base de la nominalización ‘celebración’, ‘celebrar’, es de la clase semántica A2 por lo que el argumento asociado al Grel (pronombre relativo) es arg1-pat. En el ejemplo (19), sin embargo, como el verbo base de ‘innovación’, ‘innovar’ es de la clase D1, el argumento asociado al SA argumental es arg0-agt (Véanse la Tabla 4.3 y la Tabla 4.4).

a3) Los determinantes posesivos se caracterizan por que prefieren ser interpretados como argumentos correspondientes a los sujetos verbales. También en Gurevich and Waterman (2009) se propone esta interpretación para los determinantes posesivos argumentos de nominalizaciones. Por este motivo, las reglas de

Reglas de Poss

los determinantes posesivos asignan automáticamente arg0 a este constituyente cuando especifica a nominalizaciones cuya base pertenece a las clases semánticas A o D (20), y arg1 cuando el verbo base es de la clase semántica B o C (21). Los papeles temáticos dependen de la clase semántica verbal concreta (Véase la Tabla 4.4).

(20) $[[Su]_{-arg0-agt}$ **informe**]_{SN} es correcto.

(21) Decidieron esperar $[[su]_{-arg1-tem}$ **salida**]_{SN}.

En los ejemplos anteriores, los determinantes posesivos son asociados a los diferentes argumentos y papeles temáticos porque en el primer caso (20) la nominalización ‘informe’ deriva de un verbo de la clase A2 y porque en el segundo caso (21), ‘salir’, el verbo base de ‘salida’, pertenece a la clase verbal B1 por lo que los argumentos asociados son arg0-agt y arg1-tem, respectivamente.

La Tabla 4.4 presenta la correspondencia entre las clases semánticas verbales y sus argumentos y papeles temáticos.

Clase Verbal	Sujeto Verbal	Objeto Verbal ₁	Objeto Verbal ₂
A1	arg0-cau	arg1-tem	-
A2	arg0-agt	arg1-pat	-
A31	arg0-agt	arg1-pat	arg2-loc
A32	arg0-agt	arg1-pat	arg2-ben
B1	arg1-tem	arg2-loc	-
B2	arg1-tem	arg2-efi	-
C1	arg1-tem	arg2-loc	-
C2	arg1-tem	arg2-atr	-
C3	arg1-tem	arg2-ext	-
C4	arg1-tem	arg2-ben	-
D1	arg0-agt	-	-
D2	arg0-exp	-	-
D3	arg0-scr	-	-

Tabla 4.4: Correspondencia entre la clase semántica verbal, argumentos y papeles temáticos

Reglas de dos
constituyentes

A continuación presentamos las reglas que involucran a diferentes combinaciones de constituyentes y la motivación lingüística de dichas reglas. Antes, sin embargo, me gustaría remarcar que los GRels y los Poss solo pueden aparecer una vez en el SN de núcleo deverbal y además, no pueden ser combinados puesto que aparecen en la misma posición en el SN, es decir, en la posición de especificador. En la Figura 4.2 se puede ver la frecuencia de realización en el corpus de las diferentes

4. ANOTACIÓN AUTOMÁTICA DE LOS ARGUMENTOS INTERNOS

combinaciones de constituyentes: dos SPs representan el 59 % del total, Poss y SP el 24 %, SP y SA el 14 %, Poss y SA 2 %, dos SA el 1 % y las combinaciones con GRel y SP o SA son tan poco frecuentes que no obtienen representación en la figura. Presentamos las reglas por orden de frecuencia de los constituyentes.

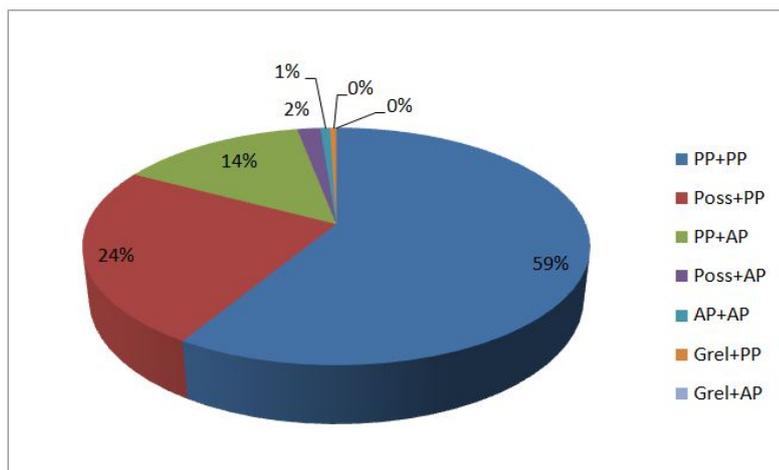


Figura 4.2: Frecuencia de las combinaciones de constituyentes en los SNs deverbales

b1) Las reglas para dos SPs (SP + SP) se resumen en la Tabla 4.5. Como en las de un SP, asumimos que un SP que complementa a una nominalización toma el argumento y papel temático del SP complemento del verbo base de la nominalización con el que comparte preposición. Sin embargo, las reglas difieren en el caso en el que no existe correspondencia entre los SPs de la nominalización y el verbo base. En estos casos, las reglas asignan arg1 al primer SP (mayoritariamente introducido por 'de') si la clase semántica del verbo base de la nominalización es A, B o C y al segundo SP se le asigna arg0 si la clase verbal es A (22) y arg2 si la clase verbal es B o C (23). En el caso de las nominalizaciones cuyos verbos base son de la clase D, las reglas asignan al primer SP un arg0 ya que es el único argumento posible en esta clase de verbos. El segundo SP se anota como un argumento adjunto (argM) sin papel temático (24). En este conjunto de reglas el orden de aparición de los constituyentes es importante puesto que determina la asignación de un argumento u otro. Los papeles temáticos asignados dependen de la subclase verbal específica a la que pertenezca el verbo base de la nominalización (Véase la Tabla 4.4).

Reglas: SP+SP

- (22) [Las **reservas** [*de oro y de divisas*]_{SP-arg1-pat} [*de Rusia*]_{SP-arg0-agt}]_{SN} subieron 800 millones de dólares.

- (23) Israel le culpa del bloqueo, en parte por sus exigencias tras [la **retirada** [de Israel]_{SP-arg1-tem} [del Líbano]_{SP-arg2-loc}]_{SN}.
- (24) Se exige el fin de la ley marcial y [el **retorno** [de un gobierno civil]_{SP-arg0-agt} [el lunes]_{SP-argM-tmp}]_{SN}.

En el ejemplo (22) tenemos un sustantivo ‘reservas’, derivado de un verbo (‘reservar’) que pertenece a la clase semántica A2 por lo que el primer SP que aparece se asocia con un arg1-pat y el segundo con un arg0-agt. Sin embargo en el ejemplo (23) al derivarse ‘retirada’ del verbo ‘retirar’, de la clase B1, el primer SP se asocia con un arg1-tem y el segundo con un arg2-loc. Si el verbo base pertenece a una clase D, como ocurre en ‘retorno’ (clase D1), el primer SP se asocia a un argumento arg0-agt y el segundo SP a un argumento adjunto (argM) que en este caso ha sido especificado como temporal (24).

	N+SP _[prep_{n1}] +SP _[prep_{n2}] y V+SP _[prep_{n1}] arg-th _{n1} +SP _[prep_n] arg-th _{n2} >N+SP-arg-th _{n1} +SP-arg-th _{n2}
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es A1 >SP ₁ -arg ₁ -tem+SP ₂ -arg ₀ -cau
Dos	N+SP ₁ [de] +PP ₂ [prep=x] si la clase semántica verbal es A2 >SP ₁ -arg ₁ -pat+SP ₂ -arg ₀ -agt
SPs	N+SP ₁ [de] +PP ₂ [prep=x] si la clase semántica verbal es A3 >SP ₁ -arg ₁ -pat+SP ₂ -arg ₀ -agt
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es B1 >SP ₁ -arg ₁ -tem+SP ₂ -arg ₂ -loc
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es B2 >SP ₁ -arg ₁ -tem+SP ₂ -arg ₂ -efi
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es C1 >SP ₁ -arg ₁ -tem+SP ₂ -arg ₂ -loc
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es C2 >SP ₁ -arg ₁ -tem+SP ₂ -arg ₂ -atr
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es C3 >SP ₁ -arg ₁ -tem+SP ₂ -arg ₂ -ext
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es C4 >SP ₁ -arg ₁ -tem+SP ₂ -arg ₂ -ben
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es D1 >SP ₁ -arg ₀ -agt+SP ₂ -argM
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es D2 >SP ₁ -arg ₀ -exp+SP ₂ -argM
	N+SP ₁ [de] +SP ₂ [prep=x] si la clase semántica verbal es D3 >SP ₁ -arg ₀ -src+SP ₂ -argM

Tabla 4.5: Notación Simplificada de las reglas específicas de dos SPs

Reglas: Poss+SP

b2) El determinante también en combinación con otros constituyentes muestra una clara preferencia por ser interpretado como el sujeto verbal. Por este motivo al SP que aparece como complemento de la nominalización en el mismo SN (el SP_{Poss}), las reglas le asignan el argumento y papel temático correspondiente al primer objeto verbal de la clase semántica del verbo base de la nominalización (Véase la Tabla 4.6). Las reglas asocian al determinante posesivo un arg0 si especifican a nominalizaciones que derivan de verbos de las clases A y D. La diferencia radica en que las nominalizaciones que derivan de la clase verbal A asignan al SP_{Poss} una interpretación de arg1 (25), mientras que si las nominalizaciones se derivan de verbos de la clase D el SP_{Poss} se anota como argM ya que el único argumento de esta clase de verbos es el arg0 (26). En SNs cuyas nominalizaciones se derivan de verbos de la clase B y C el determinante posesivo se anota como arg1 y el SP_{Poss} como arg2 (27).

4. ANOTACIÓN AUTOMÁTICA DE LOS ARGUMENTOS INTERNOS

- (25) Presentaron al juez $[[su]_{\text{Poss-arg0-agt}} \textbf{propuesta} [de \textit{solución judicial}]_{\text{SP-arg1-pat}}]_{\text{SN}}$.
- (26) $[[Su]_{\text{Poss-arg0-agt}} \textbf{paso} [por \textit{Madrid}]_{\text{SP-argM-loc}}]_{\text{SN}}$ ha dejado huella.
- (27) Justificó $[[su]_{\text{Poss-arg1-tem}} \textbf{salida} [del \textit{pais}]_{\text{SP-arg2-loc}}]_{\text{SN}}$.

En los ejemplos anteriores observamos en (25) una nominalización ‘propuesta’ cuyo verbo base (‘proponer’) pertenece a la clase semántica A2 por lo que al posesivo se le asigna la etiqueta de arg0-agt y al SP_{Poss} la de arg1-pat. En (26), en cambio, como el verbo base de ‘paso’ (‘pasar’) es de la clase D1 al posesivo también se le asigna la etiqueta de arg0-agt pero al SP_{Poss} se le asocia un argumento adjunto (argM) que en la validación manual se especifica como locativo. En (27), dado que ‘salida’ deriva de un verbo (‘salir’) de la clase B1 al posesivo se le asigna la etiqueta de arg1-tem y al SP_{Poss} la de arg2-loc.

Poss +	Poss+N+SP/SA si la clase semántica verbal es A1 >Poss-arg0-cau + SP/SA-arg1-tem
SP/SA	Poss+N+SP/SA si la clase semántica verbal es A2 >Poss-arg0-agt + SP/SA-arg1-pat
	Poss+N+SP/SA si la clase semántica verbal es A3 >Poss-arg0-agt + SP/SA-arg1-pat
	Poss+N+SP/SA si la clase semántica verbal es B1 >Poss-arg1-tem + SP/SA-arg2-loc
	Poss+N+SP/SA si la clase semántica verbal es B2 >Poss-arg1-tem + SP/SA-arg2-efi
	Poss+N+SP/SA si la clase semántica verbal es C1 >Poss-arg1-tem + SP/SA-arg2-loc
	Poss+N+SP/SA si la clase semántica verbal es C2 >Poss-arg1-tem + SP/SA-arg2-atr
	Poss+N+SP/SA si la clase semántica verbal es C3 >Poss-arg1-tem + SP/SA-arg2-ext
	Poss+N+SP/SA si la clase semántica verbal es C4 >Poss-arg1-tem + SP/SA-arg2-ben
	Poss+N+SP/SA si la clase semántica verbal es D1 >Poss-arg0-agt + SP/SA-argM
	Poss+N+SP/SA si la clase semántica verbal es D2 >Poss-arg0-exp + SP/SA-argM
	Poss+N+SP/SA si la clase semántica verbal es D3 >Poss-arg0-src + SP/SA-argM

Tabla 4.6: Notación Simplificada de las reglas específicas de Poss+SP/SA

b3) Cuando la combinación de constituyentes en un SN de núcleo de verbal son un SP y un SA, el SP parece preferir la interpretación de arg1 (en las nominalizaciones derivadas de los verbos de las clases A, B y C) cuando no hay un SP complemento del verbo con el que comparta preposición, y así lo asignan las reglas. El SA, por tanto, recibe la interpretación de arg0 en las nominalizaciones derivadas de verbos de la clase A (28) y de arg2 en las que derivan de verbos de las clases B y C. En las nominalizaciones derivadas de la clase D, las reglas priorizan como constituyente puramente argumental al SP (arg0) mientras que al SA se le asigna la etiqueta de argumento adjunto por defecto (argM). En el caso de que exista un SP complemento del verbo con el que el SP complemento nominal comparta preposición, el SP nominal toma el argumento y papel temático de dicho complemento verbal, y al SA se le asignan los argumentos correspondientes según las reglas descritas (Véase la Tabla 4.7).

Reglas: SP+SA

- (28) Estudian el desbloqueo de [las **negociaciones** [*de paz*]_{SP-arg1-pat} [*palestino-israelíes*]_{SA-arg0-agt}]_{SN}.
- (29) Un medio para lograr [una mayor **integración** [*laboral*]_{SA-arg2-loc} [*de las mujeres de la zona*]_{SP-arg1-tem}]_{SN}.
- (30) Tuvo el partido de anoche algo de [**levantamiento** [*zapatista*]_{SA-argM-mnr} [*por parte del Real Madrid*]_{SP-arg0-agt}]_{SN}.

En el ejemplo (28) observamos que en el caso de ‘negociaciones’ cuyo verbo base (‘negociar’) pertenece a la clase semántica A2 el SP recibe la etiqueta de arg1-pat y el SA la de arg0-agt. En (29), como el verbo base de ‘integración’ (‘integrar’) es de la clase B2 al SP se le asigna la etiqueta de arg1-tem y al SA la de arg2-loc. Finalmente, en (30), al ser el verbo base de la nominalización ‘levantamiento’ (‘levantar’) de la clase D1 al SP se le asigna la etiqueta de arg0-agt pero al SA se le asocia con un argumento adjunto (argM) que en la validación manual se especifica como manera. Fíjense que en las reglas que atañen a esta combinación de constituyentes el orden de aparición de los mismos no es importante; por ejemplo, en (28) el SP recibe la etiqueta de arg1 siendo el primero de los constituyentes, mientras que en (29) el SP recibe la etiqueta de arg1 siendo el segundo de los constituyentes.

SP+SA	N+SP+SA si la clase semántica verbal es A1 > SP-arg1-tem+SA-arg0-cau
	N+SP+SA si la clase semántica verbal es A2 > SP-arg1-pat+SA-arg0-agt
	N+SP+SA si la clase semántica verbal es A3 > SP-arg1-pat+SA-arg0-agt
	N+SP+SA si la clase semántica verbal es B1 > SP-arg1-tem+SA-arg2-loc
	N+SP+SA si la clase semántica verbal es B2 > SP-arg1-tem+ SA-arg2-efi
	N+SP+SA si la clase semántica verbal es C1 > SP-arg1-tem+SA-arg2-loc
	N+SP+SA si la clase semántica verbal es C2 > SP-arg1-tem+ SA-arg2-atr
	N+SP+SA si la clase semántica verbal es C3 > SP-arg1-tem+ SA-arg2-ext
	N+SP+SA si la clase semántica verbal es C4 > SP-arg1-tem+ SAP-arg2-ben
	N+SP+SA si la clase semántica verbal es D1 > SP-arg0-agt+ SA-argM
	N+SP+SA si la clase semántica verbal es D2 > SP-arg0-exp+ SA-argM
	N+SP+SA si la clase semántica verbal es D3 > SP-arg0-src+ SA-argM

Tabla 4.7: Notación Simplificada de las reglas específicas de SP + SA

Reglas: Poss+SA

b4) En el caso de la combinación de los constituyentes Poss y SA, las reglas son equivalentes a la combinación de Poss+SP. Esto es, al posesivo se le asocia el argumento correspondiente al sujeto verbal y al SA se le asigna el argumento correspondiente al primer objeto verbal de la clase semántica del verbo base de la nominalización (Véase la Tabla 4.6). Por lo tanto, si la nominalización deriva de las clases A y D, al determinante posesivo se le asigna arg0 y el SA_{Poss} se anota como arg1 (31) si la nominalización deriva de verbos de la clase A, mientras que

si las nominalizaciones se derivan de verbos de la clase D el SA_{POSS} se anota como argM (32). En SNs cuyas nominalizaciones se derivan de verbos de la clase B y C el determinante posesivo se anota como arg1 y el SA_{POSS} como arg2 (33).

- (31) [[Nuestra]-arg0-agt **experiencia** [*vital*]_{SA-arg1-pat}]_{SN} está basada en percepciones físicas.
- (32) Dio detalles de [[su]-arg0-agt **trabajo** [*diario*]_{SA-argM-loc}]_{SN}.
- (33) Se les condenó por [[sus]-arg1-tem **tendencias** [*homosexuales*]_{SA-arg2-efi}]_{SN}.

En los ejemplos anteriores se puede observar que en (31) la nominalización ‘experiencia’ cuyo verbo base, ‘experimentar’, pertenece a la clase semántica A2, tiene como especificador un posesivo al que se le asigna la etiqueta de arg0-agt y un SA_{POSS} al que se le asocia la de arg1-pat. En (32), en cambio, como el verbo base de ‘trabajo’, ‘trabajar’, es de la clase D1 al posesivo también se le asigna la etiqueta de arg0-agt pero al SA_{POSS} se le asocia con un argumento adjunto (argM) que en la validación manual se especifica como temporal. En (33), dado que ‘tendencia’ deriva de un verbo (‘tender’) de la clase B2 al posesivo se le asigna la etiqueta de arg1-tem y al SP_{POSS} la de arg2-efi.

b5) Cuando la combinación atañe a dos SAs relacionales no observamos una preferencia clara por ninguna configuración, por lo que asumimos que se comportarían de forma parecida a dos SPs (Véase la Tabla 4.8). Por lo tanto, las reglas asignan arg1 al primer SA si la clase semántica del verbo base de la nominalización es A, B o C y al segundo SA se le asigna arg0 si la clase verbal es A (34) y arg2 si la clase verbal es B o C (35). En el caso de las nominalizaciones cuyos verbos base son de la clase D, las reglas asignan al primer SA un arg0 y al segundo SA un argumento adjunto (argM) sin papel temático (36). Los papeles temáticos asignados dependen de la subclase verbal específica a la que pertenezca el verbo base de la nominalización (Véase la Tabla 4.4).

Reglas: SA+SA

- (34) [El **retoque** [*defensivo*]_{SA-arg1-pat}[*madridista*]_{SA-arg0-agt}]_{SN} funcionó bien.
- (35) El requisito para [el **estado** [*molecular*]_{SA-arg1-tem} [*metálico*]_{SA-arg2-atr}]_{SN} es que las moléculas deben estar en un estado de oxidación fraccionada.
- (36) Reclamó la celebración urgente de [una **reunión** [*ministerial*]_{SA-arg0-agt} [*europea*]_{SA-argM-adv}]_{SN}.

En el ejemplo (34) tenemos un sustantivo ‘retoque’, derivado de un verbo (‘retocar’) que pertenece a la clase semántica A2 por lo que el primer SA que aparece se asocia con un arg1-pat y el segundo con un arg0-agt, sin embargo en el ejemplo (35) al derivarse ‘estado’ de un verbo (‘estar’) de la clase C2 el primer SA se asocia con un arg1-tem y el segundo con un arg2-loc. Si el verbo base pertenece a una clase D, como ocurre en ‘reunirse’ (clase D1) (36), el primer SA se asocia

a un argumento arg0-agt y el segundo SA a un argumento adjunto (argM) que en este caso ha sido especificado como adverbial.

SA+SA	N+SA+SA si la clase semántica verbal es A1 >SA-arg1-tem+SA-arg0-cau
	N+SA+SA si la clase semántica verbal es A2 >SA-arg1-pat+SA-arg0-agt
	N+SA+SA si la clase semántica verbal es A3 >SA-arg1-pat+SA-arg0-agt
	N+SA+SA si la clase semántica verbal es B1 >SA-arg1-tem+SA-arg2-loc
	N+SA+SA si la clase semántica verbal es B2 >SA-arg1-tem+ SA-arg2-efi
	N+SA+SA si la clase semántica verbal es C1 >SA-arg1-tem+SA-arg2-loc
	N+SA+SA si la clase semántica verbal es C2 >SA-arg1-tem+ SA-arg2-atr
	N+SA+SA si la clase semántica verbal es C3 >SA-arg1-tem+ SA-arg2-ext
	N+SA+SA si la clase semántica verbal es C4 >SA-arg1-tem+ SA-arg2-ben
	N+SA+SA si la clase semántica verbal es D1 >SA-arg0-agt+ SA-argM
	N+SA+SA si la clase semántica verbal es D2 >SA-arg0-exp+ SA-argM
	N+SA+SA si la clase semántica verbal es D3 >SA-arg0-src+ SA-argM

Tabla 4.8: Notación Simplificada de las reglas específicas de dos SAs

Reglas: GRel+SP/SA

b6) En cuanto a los GRel en combinación con otro constituyente, dado que los GRel son semánticamente equivalentes a SPs introducidos por ‘de’, los anotamos como el primer SP de la combinación SP+SP, es decir, como arg1 en el caso de que la nominalización derive de las clases A, B y C y como arg0 si se deriva de la clase D. El otro constituyente se anota (**SP₂** o **SA₂**) como arg0², arg2 (37) o argM (38) según la nominalización derive de un verbo de la clase A, B y C, o D, respectivamente. Las reglas que implican a los pronombres relativos de genitivo están resumidas en la Tabla 4.9. Cabe notar que no se ha encontrado en el corpus ningún ejemplo de la combinación GRel+SA, y solo 10 ejemplos de la combinación GRel+SP.

(37) Defraudado por hombres [[cuya]-arg1-tem **permanencia** [en el equipo]_{SP-arg2-loc}]_{SN} estaba más que cuestionada.

(38) La ley de extranjería [[cuyo]-arg0-agt **paso** [por las Cortes]_{SP-argM-loc}]_{SN} ha demostrado que algunos prefieren perjudicar a su país.

En el ejemplo (37) se observa una nominalización, ‘permanencia’, que deriva de un verbo (‘permanecer’) de la clase C1, en la que el pronombre relativo se asocia con un arg1-tem y el SP con un arg2-loc. En cambio, en (38) al ser ‘paso’ una nominalización derivada de un verbo (‘pasar’) de la clase D1 al relativo se le asigna la etiqueta de arg0-agt y al SP se le etiqueta como argM, en este caso, confirmado como argM-loc.

²Ninguno de los ejemplos con relativos argumentales refleja la configuración arg1-relativo, arg0-SP.

4. ANOTACIÓN AUTOMÁTICA DE LOS ARGUMENTOS INTERNOS

GRel+	GRel+N+SP/SA si la clase semántica verbal es A1 >GRel-arg1-tem+SP/SA-arg0-cau
SP/SA	GRel+N+SP/SA si la clase semántica verbal es A2 >GRel-arg1-pat + SP/SA-arg0-agt
	GRel+N+SP/SA si la clase semántica verbal es A3 >GRel-arg1-pat + SP/SA-arg0-agt
	GRel+N+SP/SA si la clase semántica verbal es B1 >GRel-arg1-tem + SP/SA-arg2-loc
	GRel+N+SP/SA si la clase semántica verbal es B2 >GRel-arg1-tem + SP/SA-arg2-efi
	GRel+N+SP/SA si la clase semántica verbal es C1 >GRel-arg1-tem + SP/SA-arg2-loc
	GRel+N+SP/SA si la clase semántica verbal es C2 >GRel-arg1-tem + SP/SA-arg2-atr
	GRel+N+SP/SA si la clase semántica verbal es C3 >GRel-arg1-tem + SP/SA-arg2-ext
	GRel+N+SP/SA si la clase semántica verbal es C4 >GRel-arg1-tem + SP/SA-arg2-ben
	GRel+N+SP/SA si la clase semántica verbal es D1 >GRel-arg0-agt + SP/SA-argM
	GRel+N+SP/SA si la clase semántica verbal es D2 >GRel-arg0-exp + SP/SA-argM
	GRel+N+SP/SA si la clase semántica verbal es D3 >GRel-arg0-src + SP/SA-argM

Tabla 4.9: Notación Simplificada de las reglas específicas de GRel+SP/SA

En general, los SNs de núcleo deverbal con más de dos constituyentes argumentales son escasos. Esto es así porque, a diferencia de los verbos, los SNs nominalizados, al ser SNs que sirven para condensar la información, no admiten discursivamente una gran cantidad de argumentos. Esto explica que el número de argumentos es generalmente bajo: 0, 1 o 2 argumentos en la mayoría de los casos y hasta 3 o 4 argumentos en unos pocos casos (39). En estos pocos casos, si quedan constituyentes sin anotar tras la aplicación de las reglas hasta ahora descritas, existe una regla final que los anota como argumento adjunto (argM, el argumento por defecto).

Reglas de más de dos
constituyentes

- (39) Instaron a los fieles a incumplir [la **prohibición** [comunitaria]_{SA-arg0-agt} [de importar borregos de Marruecos]_{SP-arg1-pat}][a causa de un brote de fiebre aftosa]_{SP-argM-cau}]_{SN}.

En el ejemplo anterior, la nominalización ‘prohibición’ se anota por las reglas de combinación de SA+SP, pero el tercer argumento queda sin anotar por lo que la regla final lo anota como argumento adjunto (argM) por defecto, que en la validación manual se ha especificado como causa (argM-cau).

La evaluación de la eficacia de estas reglas heurísticas para la anotación de la estructura argumental de las nominalizaciones deverbales en el corpus AnCora-Es se evalúa en la siguiente sección, a partir de contrastar la anotación automática con la validación manual (descrita en el Capítulo 8).

4.2. Evaluación de la anotación automática de la estructura argumental

En esta sección se evalúa la eficacia y fiabilidad del proceso de anotación automática. Esta misma evaluación nos permite validar las hipótesis lingüísticas que subyacen a nuestras reglas heurísticas. En global, las reglas logran un 77 % de F1, calculada como media ponderada³ entre precisión y cobertura (Tabla 4.11). Este resultado demuestra que el proceso automático desarrollado para la anotación de la estructura argumental de las nominalizaciones deverbales es una estrategia válida ya que reduce el tiempo y el coste de la anotación en un 37 % si lo comparamos con una anotación completamente manual.

La Tabla 4.10 presenta los resultados obtenidos teniendo en cuenta los distintos constituyentes (en filas). La primera columna detalla cada uno de los constituyentes. La segunda columna nos informa de la frecuencia de cada uno de los constituyentes en el corpus, lo que nos da una idea de la importancia del constituyente en cuestión. La tercera y cuarta columna indican si el constituyente está anotado según una regla específica (RE) o una regla general (RG). La quinta, sexta y séptima columna presentan la Precisión (P), la Cobertura (C) y la F1 del proceso automático teniendo en cuenta todas las etiquetas asignadas. Las últimas tres columnas también muestran la Precisión, la Cobertura y la F1 del proceso automático pero esta vez teniendo en cuenta solo las etiquetas argumentales, esto es, excluyendo del cómputo la etiqueta RefMod.

Const.	Frec.	RE	RG	P	C	F1	P-RefMod	R-RefMod	F1-RefMod
Osub	11,6 %	-	+	97 %	98 %	97 %	-	-	-
Poss	6,6 %	+	-	77 %	90 %	82 %	77 %	90 %	82 %
SA	26,4 %	+	+	74 %	78 %	76 %	26 %	31 %	28 %
SAdv	0,5 %	-	+	61 %	100 %	76 %	-	-	-
SN	1,7 %	-	+	70 %	81 %	75 %	51 %	50 %	50 %
GRel	0,2 %	+	-	53 %	79 %	64 %	53 %	79 %	64 %
SP	53,0 %	+	+	51 %	53 %	52 %	51 %	53 %	52 %

Tabla 4.10: Resultados de la anotación automática por constituyentes

En la Tabla 4.11 se presentan los resultados de la anotación automática por constituyentes y etiquetas de forma detallada. En la primera columna se muestran todas las etiquetas posibles (combinación de argumentos y papeles temáticos, y RefMod) y en la segunda columna se nos indica su frecuencia. En las columnas

³En nuestra evaluación ponderamos igualmente los dos factores.

4. ANOTACIÓN AUTOMÁTICA DE LOS ARGUMENTOS INTERNOS

siguientes se especifica para cada etiqueta la F1 lograda según el tipo de constituyente. Las tres últimas columnas indican la Precisión (P), la Cobertura (C) y la F1 para cada etiqueta en global, es decir, independientemente del tipo de constituyente. Las dos últimas filas presentan los resultados para cada constituyente en global, es decir, teniendo en cuenta todas las etiquetas, y sin tener en cuenta la etiqueta RefMod.

Etiquetas	Frec.	OSub	Poss	SA	SAdv	SN	GRel	SP	P	Total	F1
		F1		C							
RefMod	36,2 %	97 %	-	89 %	76 %	83 %	-	15 %	96 %	92 %	94 %
arg0-agt	14,4 %	-	87 %	21 %	-	-	57 %	24 %	48 %	45 %	46 %
arg0-cau	0,5 %	-	23 %	0 %	-	-	-	16 %	13 %	29 %	18 %
arg0-src	0,07 %	-	100 %	-	-	-	-	90 %	83 %	91 %	87 %
arg1-Ø	0,04 %	-	-	-	-	-	-	51 %	80 %	33 %	46 %
arg1-loc	0,2 %	-	-	-	-	-	-	89 %	93 %	82 %	87 %
arg1-pat	20,7 %	-	-	31 %	-	-	66 %	66 %	52 %	75 %	61 %
arg1-tem	11,4 %	-	81 %	48 %	-	-	62 %	79 %	73 %	79 %	76 %
arg2-Ø	1,5 %	-	-	-	-	-	-	52 %	76 %	36 %	49 %
arg2-atr	0,6 %	-	-	0 %	-	-	-	19 %	35 %	8 %	13 %
arg2-ben	0,6 %	-	-	0 %	-	-	-	36 %	37 %	31 %	34 %
arg2-efi	0,2 %	-	-	28 %	-	-	-	19 %	12 %	83 %	20 %
arg2-ins	0,04 %	-	-	-	-	-	-	33 %	50 %	20 %	28 %
arg2-loc	1 %	-	-	0 %	-	-	-	29 %	28 %	29 %	29 %
arg3-ein	0,03 %	-	-	-	-	-	-	0 %	0 %	0 %	0 %
arg3-ori	0,36 %	-	-	-	-	-	-	13 %	29 %	8 %	13 %
arg4-des	0,5 %	-	-	-	-	-	-	47 %	58 %	38 %	46 %
arg4-efi	0,06 %	-	-	-	-	-	-	21 %	22 %	20 %	21 %
argM-Ø	0,9 %	-	-	6 %	-	32 %	-	6 %	3 %	58 %	6 %
argM-adv	1,4 %	-	-	-	-	-	-	11 %	22 %	7 %	11 %
argM-cau	0,5 %	-	-	-	-	-	-	8 %	38 %	4 %	8 %
argM-ext	0,6 %	-	-	-	-	-	-	0 %	0 %	0 %	0 %
argM-fin	1,6 %	-	-	-	-	-	-	66 %	57 %	73 %	64 %
argM-loc	3,7 %	-	-	-	-	50 %	-	49 %	50 %	43 %	46 %
argM-mnr	0,9 %	-	-	-	-	-	-	7 %	36 %	3 %	5 %
argM-tmp	2 %	-	-	-	-	71 %	-	58 %	78 %	44 %	56 %
Total		97 %	82 %	76 %	76 %	75 %	64 %	52 %	76 %	77 %	77 %
Total-RefMod		-	82 %	28 %	-	50 %	64 %	52 %	52 %	56 %	55 %

Tabla 4.11: Resultados de la anotación automática por constituyentes y etiquetas

Tal y como muestran las Tablas 4.10 y 4.11, cuando se tienen en cuenta todas las etiquetas posibles, los mejores resultados se logran en la anotación automática

Resultados: S y Poss

de las OSubs (97 %) y los Poss (82 %) . Los resultados tan positivos eran esperables en el primer caso puesto que las oraciones subordinadas complementos del nombre solo se consideran no argumentales y, por tanto, solo se anotan con la etiqueta RefMod. El 3 % de error se explica por casos de oraciones que son complemento de sustantivos que finalmente no se han considerado que tengan un significado de verbal. Por ejemplo, el sustantivo ‘cura’ tanto puede ser la nominalización del verbo ‘curar’ como un sustantivo sinónimo de ‘párroco’, pero solo en el primer caso se mantiene la anotación, por lo que si existen oraciones que son complementos del segundo significado de ‘cura’ son consideradas como un error. El resultado tan positivo para el determinante posesivo confirma nuestra hipótesis de que la mayoría de las ocurrencias de este constituyente se interpretan como argumentos correspondientes a los sujetos de los verbos base.

Resultados: SAs

La F1 media para los SA es de un 76 %, pero existe una diferencia significativa entre los SA no argumentales en los que la etiqueta RefMod logra un 89 % de F1 y los SA argumentales en los que en promedio se consigue solo una F1 de 28 % (Tabla 4.11). Esto implica que la regla para detectar los SA no argumentales funciona bastante mejor que la desarrollada para anotar los SA argumentales. A continuación explicamos las posibles razones para este resultado. En primer lugar, la ambigüedad de los SA relacionales respecto a su naturaleza argumental o no es un problema generalizado. Casi la mitad de los SA anotados como argumentales han sido considerados no argumentales en el proceso de validación manual, lo que implica que la hipótesis de que los SA relacionales son argumentales no siempre se verifica. De hecho, 213 lemas de nuestra lista de adjetivos relacionales (331, Apéndice A) se anotan como argumentales o no argumentales dependiendo del sustantivo al que complementan. Este fenómeno se conoce en lingüística como coocurrencia léxica. Por ejemplo, un adjetivo como ‘constitucional’ se interpreta como arg1-pat de un sustantivo como ‘reforma’, siendo el significado implícito el de la reforma de la constitución. Sin embargo, este mismo adjetivo no puede ser interpretado como argumento de un sustantivo como ‘acusación’, ni como arg1-pat ni como arg0-agt: la acusación de la Constitución (arg1-pat) o la acusación por la Constitución (arg0-agt) son interpretaciones inadecuadas. En este caso, en ‘acusación constitucional’ el adjetivo especifica el significado del sustantivo, es un tipo de ‘acusación’, acusación de que se infringe la constitución. El fenómeno de la co-ocurrencia léxica es muy frecuente entre los adjetivos relacionales que se combinan con las nominalizaciones de verbales, y explica porqué no todas las apariciones de adjetivos relacionales se comportan como argumentos. Además, existen 90 lemas adjetivales que no estaban en nuestra lista de adjetivos relacionales pero que, sin embargo, han sido anotados como argumentos. Esto se debe a la ambigüedad de algunos adjetivos. Por ejemplo, un adjetivo como ‘popular’ no se incluye en nuestra lista porque no está formado con ninguno de los sufijos detallados en la Sección 4.1.1, sin embargo cuando su significado es el de

‘relativo al pueblo’ y no el de ‘famoso, conocido’, entonces se puede interpretar como argumento. Por lo tanto, ‘popular’ en ‘movilización popular’ es arg0-agt. De esta manera, para considerar un SA argumental es más importante considerar la relación con el sustantivo al que complementa que el hecho de que el SA sea relacional o no.

Un segundo motivo que explicaría el 28 % de F1 es que el orden que asumimos en las reglas de dos SA no parece confirmarse siempre. Por un lado, el arg1 no siempre es el primer constituyente y el arg0 de la clase A y el arg2 de las clases B y C no siempre son el segundo complemento. Los resultados muestran que no existe un orden de aparición de los SAs que sea regular y muestre algún tipo de analogía con la realización de sujetos y objetos verbales. Finalmente, el mal resultado en los SA argumentales también puede explicarse porque la intuición de que los SA terceros o cuartos complementos de las nominalizaciones son argM no se confirma en los resultados, que muestran solo un 6 % de F1 (Tabla 4.11). Finalmente, existen dos tipos de errores impredecibles: (i) los que se deben a un error en la designación del sentido verbal correspondiente y (ii) los que se explican porque finalmente el sustantivo no se confirma como una nominalización deverbal.

En los SAdvs se ha conseguido un F1 del 76 %, que se traduce en un 61 % de precisión y un 100 % de cobertura (Tabla 4.11). Una precisión tan baja era de esperar puesto que ya contábamos con que algunos SAdvs se anotaran como argumentos adjuntos (argM), aunque en el proceso automático los anotamos todos como RefMod porque no había manera automática de distinguir los argMs. En el proceso de validación manual, la mayoría de los falsos positivos se han modificado como diferentes tipos de argumentos adjuntos.

Resultados: SAdvs

En los SNs que son complementos de las nominalizaciones, se logra un 75 % de F1 (Véase Tabla 4.10 y Tabla 4.11). Los SNs se pueden asociar a tres etiquetas semánticas distintas: argM-loc, argM-tmp, y la etiqueta no-argumental RefMod. Las dos primeras etiquetas se asignaban si los SNs contenían una entidad con nombre del tipo “lugar” o “fecha”, respectivamente, y la tercera se aplicaba en el caso de que el SN no contuviera ninguna entidad con nombre. Como es lógico, los mejores resultados en este constituyente son para la etiqueta RedMod—83 % de F1 (Tabla 4.11)— ya que es una regla de aplicación directa. Nuestra hipótesis de que las entidades con nombre del tipo “fecha” se corresponden con argM-tmp parece ser certera ya que se logran un 71 % de F1 (Tabla 4.11). Cabe señalar, sin embargo, que esta F1 se corresponde con una alta precisión pero una cobertura media, lo que significa que esta regla no cubre todos los casos de SNs que son argM-tmp; de hecho, casi la mitad de los SNs validados como argM-tmp se anotaron automáticamente como RefMod. Lo que nos sorprende es que la regla que asignaba argM-loc a los SNs que eran entidades con nombre del tipo “lugar” solo logre una F1 del 50 % (Tabla 4.11) porque algunos de los falsos positivos se validaron como arg1-pat.

Resultados: SNs

Resultados: GRels

La anotación automática de los GRels logra un 64 % de F1, siendo la precisión más baja (53 %) que la cobertura (79 %) (Tabla 4.10). Sin embargo, dado que el corpus solo tiene 28 ocurrencias de este tipo de constituyente dentro de un SN de núcleo deverbal, creemos que no podemos interpretar rotundamente este resultado. En general, observamos que estos pronombres no siempre realizan el arg1 de las nominalizaciones correspondientes a verbos de la clase A, B o C pero sí que realizan siempre los arg0 de las nominalizaciones correspondientes a verbos de la clase D.

Resultados: SPs

En un término medio, los **SPs** logran una F1 del 52 %, que supone un buen resultado teniendo en cuenta que a los SPs se les pueden asignar 26 etiquetas semánticas diferentes. Las razones que dan cuenta de este resultado son las siguientes: 1) el orden en la asignación de argumentos proyectado en las reglas heurísticas no siempre se mantiene, esto es, el arg1 no siempre aparece como el primer complemento, de la misma manera que el arg0 y el arg2 no siempre son el segundo de los complementos. Esto nos sugiere que no hay un orden fijo en la realización argumental de las nominalizaciones, no se puede encontrar un paralelismo en la realización de los argumentos nominales respecto al sujeto y objeto del verbo base correspondiente; 2) no todos los SPs que se realizan en segundo lugar son arg0 en las nominalizaciones derivadas de verbos de la clase A. De hecho, la gran mayoría de estos segundos SPs son argumentos adjuntos, por lo tanto, se puede decir que el arg0 no se realiza frecuentemente en los SNs de núcleo deverbal.

Aunque sabíamos que existían SPs no argumentales, en las reglas heurísticas los anotamos todos como sí fueran argumentales. Esta decisión ha sido un acierto, solo un 24 % de los SPs se han anotado como no argumentales. El resultado de los SPs también se ve afectado por el hecho de que las preposiciones de los SPs complementos verbales no son siempre las mismas que en los SPs complementos nominales. Esto queda claro en las etiquetas arg1-Ø y arg2-Ø, que son las que típicamente responden a los SPs complementos verbales. En la anotación nominal, estas etiquetas consiguen una F1 del -51 % y 52 %, respectivamente—(Tabla 4.11), pero logran una precisión muy alta y una cobertura baja. Esto indica que cuando la preposición es compartida por el SP complemento del verbo y el SP complemento del nombre, la asignación de argumento y papel temático es mayoritariamente correcta, pero que en muchas ocasiones la preposición no es compartida por lo que muchos de esos argumentos no son detectados automáticamente.

Las reglas que tienen en cuenta el tipo de preposición que introduce a un SP funcionan bastante bien (Tabla 4.12). En general se consigue un importante porcentaje de cobertura (alrededor del 90 %) pero solo porcentajes medios de precisión (alrededor del 50 %). Esto demuestra que sobregeneramos las etiquetas asignadas por estas reglas, cubriendo todos o la mayoría de los casos pero con un coste importante en cuanto a la precisión.

Preposición+Etiqu.	Precisión	Cobertura	F1
‘durante’: argM-tmp	100 %	100 %	100 %
‘tras’: argM-tmp	78 %	88 %	82 %
‘para’: argM-fin	58 %	98 %	73 %
‘sin’: argM-adv	50 %	100 %	66 %
‘según’: argM-adv	50 %	100 %	66 %
‘hacia’: arg4-des	42 %	66 %	52 %
‘desde’: arg3-ori	9 %	50 %	15 %

Tabla 4.12: Eficacia de las reglas generales para los SPs

Los resultados más sorprendentes son los de las etiquetas arg4-des y arg3-ori. En el caso de ‘hacia’, tanto la cobertura (66 %) como la precisión (42 %) son bajas, pero en el caso de ‘desde’ la cobertura es baja (50 %) y la precisión bajísima (9 %). La razón de esto es que el 40 % de los SPs introducidos por ‘desde’ han sido reanotados como argM-tmp. Nuestra intuición sobre la relación semántica entre dicha preposición y el concepto de origen no se ha confirmado. De hecho, si hubiésemos asociado la preposición ‘desde’ con un argumento temporal (argM-tmp), la cobertura hubiese sido del 100 % y la precisión del 36 %, lo que se traduciría en un F1 del 53 %, en lugar de la F1 del 15 % que se obtiene asociando la preposición ‘desde’ con el concepto de origen. Con este cambio, el resultado global de la anotación automática se incrementaría hasta el 78 % de F1.

Además, existen dos tipos de errores que explican también los resultados obtenidos en los SPs: los errores parciales y los incontrolables. Con errores parciales nos referimos a aquellos casos en los que solo hay un cambio en la etiqueta semántica asignada, bien la posición argumental o bien el papel temático. Por ejemplo, el 19 % de los argM-loc asignados por la regla de entidad con nombre “lugar” se reanotaron como arg1-loc o arg2-loc. De manera similar, la etiqueta por defecto argM (sin papel temático) para los terceros y cuartos argumentos nominales solo ha necesitado la especificación del papel temático en el 40 % de los casos. En el porcentaje restante la etiqueta se ha modificado completamente. Por lo tanto, en los SPs sí se puede confirmar parcialmente que los terceros y cuartos argumentos nominales son argM. Los errores impredecibles son de tres tipos: (i) los que se explican porque finalmente el sustantivo no se confirma como una nominalización deverbal (2 %); (ii) los que se corresponden con anotaciones automáticas que se han corregido con etiquetas erróneas (6 %); y (iii) los que se deben a un error en la designación del sentido verbal correspondiente.

Si la etiqueta RefMod no se incluyera en la evaluación el resultado global descendería en un 22 % (Véase Tabla 4.11). Esto es así porque en los constituyentes en los que se aplica esta etiqueta (OSub, SAdv, SN y SA), las reglas para detec-

Resultados: RefMod

tar los RefMod funcionan muy bien. Por ejemplo, en el caso de las OSubs y los SAdvS, como es la única etiqueta posible los resultados son muy buenos (97 % y 76 % respectivamente). Por lo tanto, si no tenemos en cuenta estos dos la corrección global del proceso de anotación automática disminuiría. De la misma manera, en los SNs y SAs, pese a no ser la única etiqueta posible, RefMod es la etiqueta en la que la asignación automática funciona mejor. Recuérdese que en los SAs la etiqueta RefMod conseguía una F1 del 89 %, mientras que en la asignación de los SAs argumentales solo se conseguía el 28 %. En los SNs, aunque la diferencia entre la etiqueta RefMod y los SNs argumentales no es tan importante, no es menos cierto que en la asignación de los SNs argumentales se consigue un 50 % comparado con el 83 % (Tabla 4.11) conseguido en la asignación de la etiqueta RefMod en los SNs. Por lo tanto, si no se tiene en cuenta esta etiqueta, desciende la F1 en la anotación de estos constituyentes y en consecuencia en la globalidad de la anotación automática. Una vez dicho esto, queremos señalar que para nosotros es muy importante que esta etiqueta se tenga en cuenta en la evaluación: por una parte, desde un punto de vista estrictamente lingüístico, es necesario tener identificados a los constituyentes que no pueden ser argumentos de las nominalizaciones y este es nuestro objetivo con la etiqueta RefMod. Por lo tanto, las reglas heurísticas automáticas han ahorrado mucho tiempo en la anotación del corpus, que era nuestro principal objetivo. Por otra parte, si comparamos nuestros resultados con los de sistemas nominales de etiquetado semántico, se puede decir que las reglas para identificar los constituyentes RefMod son similares a la tarea de identificación de argumentos en estos sistemas, es decir, en la tarea de identificar qué constituyentes son argumentos y cuáles no. Dado que nuestras reglas RefMod identifican los constituyentes no argumentales, creemos firmemente que deben formar parte de la evaluación, de la misma manera que forma parte la tarea de identificación de argumentos en la evaluación de los sistemas nominales de etiquetado semántico.

En la siguiente sección se discuten los principales hallazgos obtenidos a partir de la anotación automática descrita aquí. También se compara los resultados de este trabajo con el de otros sistemas de etiquetado semántico nominal desarrollados básicamente para el inglés.

4.3. Discusión

En las secciones anteriores se ha descrito y evaluado el conjunto de reglas heurísticas que nos permite anotar automáticamente la estructura argumental de las nominalizaciones deverbales en el corpus AnCora-Es. La hipótesis inicial que subyace a este trabajo de que las nominalizaciones deverbales heredan la estructura argumental del verbo correspondiente parece confirmarse puesto que este conjunto de reglas, que básicamente se basan en la información codificada en el léxico

AnCora-Verb, logra un rendimiento global del 77 % F1 (si tenemos en cuenta la etiqueta RefMod). Sin embargo, no todas las hipótesis lingüísticas que subyacen a las distintas reglas se confirman en el mismo grado. A continuación detallamos qué reglas funcionan mejor, y por lo tanto, qué hipótesis lingüísticas se corroboran.

En lo que respecta a las reglas generales, se puede decir que funcionan bastante bien. Las reglas para la detección de los complementos RefMod logran una F1 del 94 % (Tabla 4.11), lo que significa que la hipótesis que mantiene que los SAs no relacionales, los SAdvS, los SNs y las Ss no son argumentos de nominalizaciones (Badia, 2002; Meyers, 2007; Picallo, 1999) se confirma. Sin embargo, para los SNs nuestra propuesta es que aquellos que contengan una entidad con nombre del tipo “lugar” o “fecha” son argumentos locativos (argM-loc) o temporales (argM-tmp), respectivamente. Esto se confirma parcialmente para la asignación de la etiqueta argM-loc (50 % F1) y ampliamente para la asignación de la etiqueta argM-tmp (71 % F1) en los SNs (Tabla 4.11). Las reglas que tienen en cuenta el tipo de preposición que introduce a un SP también logran un buen resultado (Véase la Tabla 4.12), lo que confirma que ciertas preposiciones apuntan a un tipo específico de argumento adjunto.

Uno de los principales resultados que se pueden extraer de la actuación de las reglas específicas es que no existe un orden fijo en la realización nominal de los argumentos que se corresponden con los argumentos verbales, es decir, que no se puede afirmar que los arg1 de los verbos se realizan siempre en primer lugar en una configuración nominal. En los SPs y los SAs, el arg1 no siempre se realiza como el primer complemento y el arg0 (para la clase A) y el arg2 (clases B y C) no siempre aparecen como el segundo complemento. Sin embargo, el orden inverso, es decir, el arg1 asociado al segundo complemento, tampoco hubiera conseguido mejores resultados. Los casos analizados del corpus muestran que el orden de los constituyentes en el dominio nominal es más libre que en el dominio verbal y en cierta medida depende del contexto. Por eso mismo, un patrón sintáctico semántico específico nunca es lo suficientemente abarcador ya que es el contexto el que proporciona la información necesaria para asociar al arg1 con el primer o segundo constituyente en un SN de núcleo deverbal. Además este orden más libre de los argumentos también se ve motivado por un mayor grado de opcionalidad de los argumentos. De hecho, observamos que el arg0 es un argumento opcional que casi nunca se realiza en los SNs de núcleo deverbal, lo que constituye una observación muy interesante desde el punto de vista lingüístico.

Concretamente en los SPs, vale la pena mencionar que las preposiciones regidas no siempre son compartidas por los complementos preposicionales verbales y los nominales: las etiquetas arg1-Ø y arg2-Ø son las que típicamente responden a los SPs complementos verbales. Las reglas para su anotación logran una precisión muy alta y una cobertura baja. Esto indica que cuando la preposición del SP

complemento verbal es compartida por el SP complemento nominal, la asignación de argumento y papel temático es mayoritariamente correcta, pero que en muchas ocasiones la preposición no es compartida por lo que muchos de esos argumentos no son detectados automáticamente.

En el caso de los SAs, se ha demostrado que algunos adjetivos relacionales (45 %) no son argumentales, poniendo en duda la hipótesis basada en la bibliografía (Picallo, 1999) y (Grimshaw, 1990) de que este tipo de constituyente son siempre argumentales. Lo que emerge de este análisis es que los adjetivos relacionales están sujetos al fenómeno de la co-ocurrencia léxica, es decir, son argumentales dependiendo del sustantivo al que complementan.

En cuanto a los determinantes posesivos, queda confirmada con un 82 % de F1 nuestra hipótesis inicial de que este tipo de constituyente se interpreta mayoritariamente como el argumento correspondiente al sujeto de los verbos base correspondientes, hipótesis compartida para el inglés con Gurevich and Waterman (2009). Respecto a los pronombres relativos de genitivo, debido a su escasez en el corpus (solo 28 ocurrencias) creemos que no se puede extraer ninguna conclusión definitiva. Finalmente, la regla por defecto de asignar argM al tercer o cuarto SA o SP tiene buen resultado en el segundo de los constituyentes pero no en el primero, que se han corregido mayoritariamente como RefMod. Esto confirma que en español los SPs tienden a ser argumentos de las nominalizaciones mientras que los SA suelen ser simples modificadores.

4.3.1. Comparación de resultados

En esta subsección comparamos los resultados que hemos obtenido con los de aquellos trabajos presentados en el Capítulo 2 que se ocupan de la anotación de los argumentos de las nominalizaciones principalmente para el inglés, siendo este trabajo la primera aproximación que se centra en la estructura argumental de las nominalizaciones del español.

Hull and Gomez, 2000

El sistema propuesto por Hull and Gomez (2000) logra unos resultados muy buenos en la interpretación de los genitivos (93 % de corrección), los SPs (96 %) y los SAs (71 %). Sin embargo, solo se llevó a cabo en un subconjunto de 10 nominalizaciones diferentes (1.247 ocurrencias). Es una muestra muy pequeña que no proporciona una idea clara de cómo actuarían sus algoritmos en un conjunto de nominalizaciones más amplio. Por lo tanto, parece difícil comparar estos resultados con los que presentamos en este trabajo, que se han obtenido a partir de la validación manual de 1.655 lemas diferentes de nominalizaciones deverbales, correspondientes a 23.431 ocurrencias.

CoNLL-2008

NomBank se ha usado como corpus de entrenamiento para sistemas de etiquetado semántico nominal basado en técnicas de aprendizaje automático supervisado como son los trabajos de Che et al. (2008), Johansson and Nugues (2008), Zhao

and Kit (2008) and Ciaramita et al. (2008) presentados en la CoNLL-2008 Shared Task en *Joint Parsing of Syntactic and Semantic Dependencies* (Surdeanu et al., 2008)⁴. En la tarea de asignar argumentos (papeles temáticos solo en el caso de los argumentos adjuntos) a las nominalizaciones deverbales, el mejor resultado es el logrado por Che et al. (2008) que obtiene una F1 de 76,64 %. En este caso, los participantes tenían 20 etiquetas diferentes que asignar, mientras que en nuestro sistema las etiquetas posibles a asignar son 26 ya que también tenemos en cuenta los roles semánticos en los argumentos nucleares.

Otra manera de realizar la tarea de etiquetado semántico nominal es la aproximación no supervisada presentada por Padó et al. (2008). En este trabajo se parte de las anotaciones verbales de FrameNet para llevar a cabo la tarea de etiquetado semántico en las nominalizaciones deverbales. Un modelo híbrido que combina información sintáctica con rasgos semánticos distribucionales logra el mejor resultado (56,42 % F1). Si se tiene en cuenta que este modelo no aprende sobre anotación nominal sino verbal y que FrameNet tiene unos roles semánticos más finos que NomBank, creemos que este resultado es muy bueno, comparable a nuestra eficiencia global sin tener en cuenta la etiqueta RefMod (Tabla 4.11). Sin embargo, el problema con los sistemas basados en técnicas de aprendizaje automático es que no proporcionan evaluaciones específicas para tipos de constituyentes o etiquetas semánticas y, por lo tanto, no se pueden extraer de ellos observaciones lingüísticas. La única forma de comparación es la actuación global de los sistemas y aquí, nuestras reglas heurísticas superan los sistemas de etiquetado semántico automático.

Padó et al., 2008

El trabajo más similar al nuestro es el de Gurevich and Waterman (2009) que también han diseñado un serie de reglas heurísticas para anotar las nominalizaciones deverbales a partir de un léxico verbal. Disponen de dos muestras para el test que son también bastante grandes, una de dos millones de documentos y otra de un subconjunto de 10.000 documentos (de esta manera evalúan también si el tamaño de la muestra tiene incidencia en el resultado). Sin embargo, existen dos diferencias importantes: 1) en su trabajo analizan si los complementos de las nominalizaciones son interpretables como ‘casi-sujetos’ (+Subj) o ‘casi-objetos’ (+Obj), esto es, solo asignan dos etiquetas posibles, mientras que en este trabajo contemplamos 26 etiquetas semánticas posibles, lo que supone una mayor dificultad respecto a las dos etiquetas (+Subj, +Obj) empleadas por Gurevich y Waterman. 2) Estos autores solo trabajan con SPs introducidos por la preposición inglesa ‘of’, *de* (‘of’-PPs) y con determinantes posesivos (Poss), mientras nosotros trabajamos con cualquier constituyente que ocurra en un SN de núcleo deverbal (SPs, SAs, GRel, Poss, SNs, SAdvS, y OSubs).

Gurevich and Waterman,
2009

Por lo tanto, para comparar nuestros resultados con los suyos nos centramos

⁴<http://www.clips.ua.ac.be/conll2008/>

en los SPs y los Poss. La F1 lograda por Gurevich and Waterman (2009) en la anotación de ‘of’-PPs es de un 82 %. Esta cifra supera en un 30 % la F1 lograda por nuestras reglas heurísticas en este constituyente. Sin embargo, tres importantes factores se tienen que tener en cuenta: en primer lugar, nuestras reglas de SPs incluyen todos los tipos de SPs, tienen en cuenta todas las preposiciones posibles y no solo *de*-SPs— equivalente español a los ‘of’-PP. En segundo lugar, el ya comentado aumento de la dificultad que supone la asignación de 26 etiquetas frente a solo dos etiquetas. Finalmente, nuestra evaluación incluye tanto a los SPs que son los únicos constituyentes en el SN así como a los SPs que aparecen en combinación con otros constituyentes mientras que en su trabajo solo evalúan los SPs en el primer caso. Estos tres factores influyen en la diferencia de F1 entre ambos trabajos. En lo que respecta a los determinantes posesivos, su trabajo logra un 85 % de F1, lo que supone una mejora del 3 % sobre nuestro resultado en este constituyente. Sin embargo, cabe notar que 1) nuestras reglas de Poss asignan un mayor número de etiquetas con el aumento de dificultad que eso supone y 2) nuestra evaluación incluye tanto a los Poss que son los únicos constituyentes en el SN así como a los Poss que aparecen en combinación con otros constituyentes mientras que en su trabajo solo se evalúan los Poss del primer caso. Estos dos factores nos llevan a concluir que nuestro 82 % es, en efecto, un mejor resultado que su 85 %.

4.4. Conclusiones

En este capítulo se ha presentado el paquete de reglas heurísticas RHN, que ha permitido la anotación automática de la estructura argumental de las nominalizaciones deverbales con un 77 % de corrección en F1. La evaluación de RHN también ha permitido la evaluación de las hipótesis lingüísticas que subyacían en él. La principal hipótesis lingüística, la que asume que las nominalizaciones deverbales heredan su estructura argumental de los correspondientes verbos, se ha visto ampliamente secundada por el buen resultado de la anotación automática. Además, también nos ha permitido observar importantes resultados sobre la realización argumental de las nominalizaciones deverbales (especialmente acerca de los constituyentes y el orden de los mismos). El buen funcionamiento de este conjunto de reglas heurísticas parece indicar que sería transportable a lenguas románicas cercanas al español como el catalán, el italiano o el francés.

Parte III
Denotación

CAPÍTULO 5

LA DENOTACIÓN EN LAS NOMINALIZACIONES DEVERBALES: ESTUDIO EMPÍRICO

En este capítulo se presenta el estudio realizado para analizar la distinción denotativa entre evento y resultado. Este estudio consta de dos partes: la primera consiste en el análisis lingüístico del comportamiento morfosintáctico y semántico de los sustantivos deverbales atendiendo especialmente a los criterios aceptados en la bibliografía para establecer la distinción entre evento y resultado (descritos en el Capítulo 2) con el objetivo de esclarecer si dichos criterios, la mayoría aplicados al inglés, son válidos para el español (Sección 5.1). Este análisis lingüístico, además, nos ha permitido encontrar otras pruebas o características para identificar la denotación de las nominalizaciones deverbales útiles de forma complementaria a los criterios presentados en la bibliografía. En segundo lugar, y para cerrar el análisis empírico de esta distinción denotativa, se describen una serie de experimentos basados en técnicas de aprendizaje automático que nos han permitido evaluar, positivamente, la consistencia de los atributos considerados pertinentes para establecer la distinción denotativa y detectar los rasgos más relevantes para dicha distinción (Sección 5.2). Finalmente, presentamos nuestras conclusiones en la Sección 5.3.

5.1. Denotación: análisis lingüístico

El estudio empírico de la denotación se realizó al mismo tiempo que el de la estructura argumental, de ahí que la muestra de datos sea la misma: las 3.077 ocurrencias correspondientes a los 817 lemas de nominalizaciones deverbales de un subconjunto de 100.000 palabras del corpus AnCora-Es (véase la Sección 3.1 para

ver cómo se ha seleccionado esta muestra de datos). El análisis lingüístico consiste en observar estas nominalizaciones en su contexto y a partir de ahí clasificarlas en eventos o resultados (Peris and Taulé, 2009). Este análisis estuvo enfocado a la reflexión y a la obtención de datos que nos permitieran obtener una caracterización de la denotación en las nominalizaciones. La clasificación se llevó a cabo por dos expertos lingüistas que en todo momento podían comparar y comentar su decisión sobre el tipo denotativo de las nominalizaciones y en todos los casos las decisiones eran acordadas. Durante el proceso de clasificación, se observó que no siempre era posible distinguir entre evento y resultado porque el contexto, es decir, la oración en la que aparece la nominalización, no era suficientemente informativo, por lo que establecimos una nueva categoría a la que denominamos “subespecificado”. También observamos que las nominalizaciones aparecían en numerosas ocasiones en construcciones más amplias que están lexicalizadas, es decir, que constituyen una expresión idiomática como por ejemplo ‘centro de atención’. Distinguímos seis tipos de construcciones lexicalizadas en función de su similitud con diferentes categorías morfológicas: lexía nominal (‘síndrome de abstinencia’), verbal (‘estar de acuerdo’), adjetival (‘al alza’), adverbial (‘con cuidado’), preposicional (‘en busca de’) o conjuntiva (‘en la medida que’). Solo las lexías nominales fueron asociadas con uno de los tres tipos denotativos propuestos –evento, resultado, subespecificado–. Es importante distinguir entre los diferentes tipos de construcciones lexicalizadas puesto que si se trata de una lexía no nominal, entonces no recibe denotación alguna ya que es una distinción semántica solo asociada a sustantivos.

Así pues, teniendo en cuenta 1) la clasificación semántica realizada, 2) la anotación morfológica, sintáctica y semántica previamente codificada en AnCora-Es y 3) la información codificada en el léxico verbal AnCora-Verb, podemos contrastar la validez de los criterios propuestos en la bibliografía (véase el Capítulo 2) para establecer la distinción de evento y resultado. Cabe puntualizar aquí que un mismo lema puede tener sentidos distintos, es decir, puede estar asociado a denotaciones distintas, lo que se refleja en un comportamiento morfosintáctico diferenciado. Por esta razón, a partir de ahora hablaremos de sentidos nominales. Las 3.077 ocurrencias se clasificaron finalmente en 1.121 sentidos. De estos 1.121 sentidos, 807 fueron anotados como resultados (72 %), 113 como eventos (10 %), 131 como subespecificados (12 %) y 70 como lexías no-nominales (6 %).

5.1.1. Análisis de los criterios de la bibliografía

Los criterios lingüísticos propuestos para distinguir entre nominalizaciones eventivas y resultativas, como vimos en la Sección 2.1.1, son de distinta naturaleza: algunos hacen referencia a cuestiones morfosintácticas (pluralización, tipo de determinante) y otros, en cambio, a cuestiones sintáctico-semánticas (obligatoriedad del argumento interno, verbo del que deriva, etc.). El corpus AnCora-Es, anotado a diferentes niveles lingüísticos, nos permitió obtener la información morfológica y sintáctico-semántica de los SNs cuyos núcleos son los sustantivos deverbales extraídos, es decir, las características morfológicas del sustantivo y las características sintáctico-semánticas de sus complementos. Además de la información lingüística explícita en el corpus, utilizamos también el lexicón AnCora-Verb, del cual obtuvimos información sobre la clase semántica de los verbos de los que derivan los sustantivos analizados.

De los doce criterios seleccionados de las propuestas de los distintos autores, que se reproducen en la Tabla 2.2 (Sección 2.1.1), los seis primeros (clase verbal de la que deriva el sustantivo, pluralización, tipo de determinante, preposición que introduce al complemento agente, obligatoriedad del argumento interno y distinción entre argumentos externos y poseedores) resultaron más fáciles de evaluar porque los datos que requieren se encontraban con relativa facilidad en el subconjunto de 100.000 palabras seleccionado del corpus AnCora-Es. Sin embargo, los criterios restantes eran más difíciles de evaluar, bien porque los ejemplos que tenemos son tan escasos que no se podían obtener resultados reveladores sobre ellos (predicado verbal con el que combina la nominalización, modificadores aspectuales y estructuras de control), bien porque no se encontró ningún ejemplo en el corpus (modificadores del agente), o bien porque ni en el lexicón AnCora-Verb ni en el corpus AnCora-Es se disponía de información codificada sobre ellos (afectación del objeto, telicidad del verbo base). De estos criterios, finalmente solo se pudo evaluar el predicado verbal con el que se combina la nominalización, y para ello fue necesario ampliar la muestra analizada al total de palabras del corpus AnCora-Es (500.000 palabras) para poder obtener datos suficientes para su evaluación. Creemos que se trata de un criterio interesante porque relaciona la denotación del sustantivo con el tipo de predicado verbal con el que combina, superando los límites del SN, a diferencia del resto de criterios.

Evaluación de los
criterios

A continuación detallamos cómo se aplicaron y contrastaron cada uno de los criterios evaluados y se muestran los resultados para aquellos que nos fue posible obtener datos suficientemente significativos (Véase la Tabla 5.1). Recuérdese que la descripción de cada uno de los criterios se halla en el Capítulo 2, en la Sección 2.1.

Clase Verbal

1. Clase Verbal (fila 1 de la Tabla 5.1). Para aplicar y analizar este criterio partimos de la clasificación semántica de verbos propuesta en AnCora-Verb. Como vimos en la Sección 4.1.1, en este lexicón, cada predicado verbal se relaciona con una o más clases semánticas en función esencialmente de los cuatro tipos básicos de eventos que denotan siguiendo la propuesta de Vendler (1967) -realizaciones (clase semántica A), logros (clase semántica B), estados (clase semántica C) y actividades (clase semántica D)- y de las alternancias de diátesis en las que el verbo participa (causativa-incoativa, activa-pasiva, etc). Las realizaciones se corresponden con predicados transitivos, los logros se vinculan en general con los verbos inacusativos, los estados se relacionan con los verbos estativos y las actividades se corresponden con los verbos inergativos. De esta manera, utilizamos el léxico AnCora-Verb como referencia para consultar las clases verbales de los verbos a partir de los cuales derivan los 817 sustantivos que conforman nuestra muestra de análisis. Esto nos permite examinar si las afirmaciones acerca de la relación entre las denotaciones de los sustantivos deverbales y las clases verbales de los verbos correspondientes se mantienen en los datos del español que se analizan.

En la muestra analizada, la mayoría de los sentidos nominales son resultativos (72 %) por lo que no es de extrañar que todas las clases verbales tengan un mayor porcentaje de este tipo de denotación entre los sustantivos que derivan. Sin embargo, lo que es realmente significativo es que los verbos estativos (clase semántica C) y los que denotan actividades, básicamente inergativos-intransitivos (clase D) dan lugar casi exclusivamente a sustantivos deverbales resultativos, 97 % y 100 % respectivamente. Mientras que los verbos que denotan realizaciones (verbos transitivos, clase A) y logros (verbo inacusativos, clase B) admiten tanto una lectura eventiva, resultativa o subespecificada, confirmando la hipótesis de Picallo (1999). Cabe destacar también que los sustantivos eventivos se derivan mayoritariamente de verbos transitivos (15 % frente al 1 % de sustantivos eventivos derivados de verbos inacusativos) y que los sustantivos subespecificados lo hacen en su mayoría de verbos inacusativos (28 % frente al 11 % de sustantivos subespecificados derivados de verbos transitivos y el 3 % de verbos estativos).

Pluralización

2. Capacidad de pluralización (fila 2 de la Tabla 5.1). La capacidad de pluralización de los 817 sustantivos se ha medido teniendo en cuenta si aparecen o no en plural en alguna ocurrencia de la muestra analizada. Los resultados obtenidos –98 % de las nominalizaciones que pueden aparecer en plural se clasificaron como resultativas y el restante 2 % como subespecificadas –confirma que la pluralidad es una característica que mayoritariamente indentifica a los sustantivos resultativos. El singular, en cambio, no es un rasgo decisivo para descartar ninguna de las denotaciones nominales y su distribución es paralela a la de los sentidos nominales; esto es, hay más sustantivos resultativos en singular porque en general existen más sentidos nominales resultativos –69 % de las nominalizaciones en singular

fueron clasificadas como resultados, 15 % como eventos y un 16 % como subespecificados. No obstante, es importante señalar que los sustantivos eventivos en su totalidad y los subespecificados en gran parte, aparecen únicamente en singular. A pesar de estos resultados, cabe señalar que es semánticamente posible encontrar una nominalización en plural que tenga una lectura eventiva. Por ejemplo, en (1) ‘bombardeos’ se refiere a las múltiples acciones de bombardear, por lo que se da la lectura eventiva.

(1) [Los **bombardeos**_{<evento>} de Sarajevo por parte del ejército serbio]_{SN}.

3. Tipo de determinante (fila 3 de la Tabla 5.1). Para aplicar y evaluar este criterio hemos tenido en cuenta el tipo de determinante que aparece en la posición de especificador de los SNs cuyos núcleos son los sustantivos deverbales analizados. Los determinantes que pueden ocupar esta posición son los determinantes definidos, los indefinidos, los demostrativos, los posesivos y los numerales o bien puede estar vacía. Como ocurría anteriormente, dado que los sentidos nominales resultativos son los más abundantes, son también los que mayor porcentaje de cada tipo de determinante tienen. Sin embargo, el dato significativo aquí es que los determinantes indefinidos (99 %), los demostrativos (100 %) y los cuantificadores (100 %) aparecen de manera casi exclusiva con aquellos sentidos nominales clasificados como resultativos. El determinante definido, el posesivo y la posición vacía del especificador pueden ocurrir en las tres clases nominales. El 72 % de los determinantes definidos son especificadores de nominalizaciones de resultado, el 13 % de nominalizaciones de evento y el 15 % de nominalizaciones subespecificadas. El 82 % de los determinantes posesivos ocurren con nominalizaciones de resultado, el 10 % con nominalizaciones de evento y el 8 % con nominalizaciones subespecificadas. Finalmente, el 88 % de las nominalizaciones sin especificación son clasificadas como resultados, el 5 % como eventos y el 7 % como subespecificadas. Los datos, por tanto, confirman solo de manera parcial las hipótesis de partida: si bien es cierto que los sustantivos resultativos aparecen con una gama más amplia de determinantes, no parece confirmarse que los sustantivos eventivos tengan que ser especificados siempre con el determinante definido ya que el posesivo y la opción de no determinante también son posibles según los datos.

Determinante

4. Preposición + Agente (fila 4 de la Tabla 5.1). Para aplicar y contrastar este criterio se han considerado los complementos agentivos introducidos por ambos tipos de preposición que efectivamente aparecen en la muestra de corpus analizada. Los sintagmas preposicionales (SPs) que se interpretan como agente en los SNs analizados están introducidos por las siguientes preposiciones: ‘de’, ‘entre’, ‘por’ y ‘por parte de’. La distribución de los cuatro tipos de SPs es complemen-

Preposición+Agente

taria entre las dos denotaciones (evento y resultado): el complemento nominal agentivo introducido por ‘de’ o ‘entre’ aparece en sustantivos resultativos (98 % y 100 %, respectivamente). Sin embargo, cuando la preposición que introduce al complemento agentivo es ‘por’ o ‘por parte de’ la lectura del sustantivo es eventiva (100 % en ambos casos). En este sentido, la hipótesis inicial parece corroborarse y los diferentes tipos de SPs pueden ser un buen indicador de la interpretación del sustantivo deverbal.

Argumento Interno

5. Obligatoriedad del argumento interno (fila 5 de la Tabla 5.1). Para aplicar y analizar este criterio se tuvo en cuenta aquellos nominales en los que el argumento interno está sintácticamente explícito y el tipo de constituyente que lo realiza. Como resultado observamos que la mayoría de las nominalizaciones eventivas estaban complementadas por un argumento interno (98 %). Esto también ocurría con las nominalizaciones clasificadas como subespecificadas en un porcentaje bastante amplio (78 %). Sin embargo, este porcentaje descendía considerablemente en el caso de las nominalizaciones de resultado (34 %). Estos datos, por tanto, confirman las hipótesis de Picallo (1999) y Grimshaw (1990).

En la Tabla 5.1 se muestra que son cuatro los constituyentes en que se realiza el argumento interno: los determinantes posesivos, los SPs, los pronombres relativos y los adjetivos relacionales. Los dos primeros se caracterizan por aparecer en las tres clases nominales: el 41 % de los determinantes posesivos aparecen con nominalizaciones de resultado, el 38 % con nominalizaciones de evento y el 21 % con nominalizaciones subespecificadas; y el 53 % de los SPs complementan a nominalizaciones de resultado, el 25 % a nominalizaciones de evento y el 22 % a nominalizaciones subespecificadas. Los pronombres relativos aparecen solo en nominales eventivos (29 %) y resultativos (71 %), mientras que los adjetivos relacionales ocurren exclusivamente en nominales resultativos (97 %), lo que constituye una marca de identificación de estos sustantivos, como afirma Picallo (Véase el siguiente criterio).

Poseedores vs. argumentos

6. Poseedores vs. argumentos (fila 6 de la Tabla 5.1). Para aplicar y analizar este criterio se han tenido en cuenta las ocurrencias de los tres constituyentes que están en juego en este criterio (adjetivos relacionales, determinantes posesivos y SPs introducidos por la preposición ‘por’¹) y hemos observado si se interpretan como argumentos externos y si esto condiciona la denotación del nominal. Para decidir si estos constituyentes son argumentos externos hemos parafraseado el SN con su estructura oracional correspondiente; si estos constituyentes son equivalentes al argumento externo del verbo, se han considerado argumentos externos. Por ejemplo, en (2) ‘la sociedad’ es argumento externo de ‘se manifestó’, si la no-

¹El equivalente español a las *by-phrases* del inglés.

5. LA DENOTACIÓN EN LAS NOMINALIZACIONES DEVERBALES: ESTUDIO
EMPÍRICO

Criterios	Valores	R	E	SE
Clase Verbal	Realizaciones	74 %	15 %	11 %
	Logros	71 %	1 %	28 %
	Estados	97 %	-	3 %
	Actividades	100 %	-	-
Pluralización	Plural	98 %	-	2 %
	Singular	69 %	15 %	16 %
Determinantes	Definido	72 %	13 %	15 %
	Indefinido	99 %	-	1 %
	Demostrativo	100 %	-	-
	Posesivo	82 %	10 %	8 %
	Cuantificador	100 %	-	-
	Sin determinante	88 %	5 %	7 %
Preposición + Agente	de	98 %	-	2 %
	entre	100 %	-	-
	por	-	100 %	-
	por parte de	-	100 %	-
Argumento interno	Posesivo	41 %	38 %	21 %
	SPs	53 %	25 %	22 %
	Pronombre relativo	71 %	29 %	-
	Adjetivo Relacional	97 %	-	3 %
Argumento externo	por-PPs	-	100 %	-
	SA Relacional	100 %	-	-
	Poss	95 %	-	5 %
Predicados	Atributivo	75 %	6 %	18 %
	Eventivo	44 %	41 %	15 %

Tabla 5.1: Resultados de los criterios por denotaciones. Leyenda: R= resultado; E= evento; y SE= subespecificado.

minimalización de verbal correspondiente, ‘manifestación’ en (3), tiene algún complemento equivalente a ‘la sociedad’, como es el adjetivo ‘social’, entonces este complemento se considera argumento externo.

- (2) La sociedad **se manifestó** mucho más en la década de los 80.
- (3) [Las **manifestaciones** sociales]_{SN} aumentaron en la década de los 80.

Como se ha comentado anteriormente, en este criterio no había acuerdo entre los

distintos autores sobre si los SAs relacionales, los Poss y los SPs introducidos por la preposición ‘por’ se podían interpretar como argumento externo. Los resultados que se han obtenido son claros: los SPs introducidos por la preposición ‘por’ con interpretación de argumento externo son los únicos que ocurren en SNs cuyos núcleos son sustantivos eventivos. Los SAs relacionales, por su parte, interpretados como argumentos externos aparecen de manera exclusiva en SNs cuyos núcleos son resultativos. Los Poss con interpretación de argumentos externos reparten sus apariciones entre sustantivos clasificados como resultativos (95 %) o como subespecificados (5 %), si bien entre los primeros son notablemente más abundantes. Así pues, la hipótesis de Grimshaw (1990) se confirma parcialmente para el español ya que solo los SPs introducidos por la preposición ‘por’ son garantes de una lectura eventiva en español. Respecto a los adjetivos relacionales, es la tesis de Picallo (1999) la que parece verificarse puesto que los datos muestran que los adjetivos relacionales solo aparecen como argumentos en los nominales resultativos. Además, se observa una tendencia de los determinantes posesivos a realizarse como argumentos externos predominantemente en nominales resultativos, lo que no confirma ninguna de las propuestas teóricas (Picallo afirmaba que podían aparecer también en nominales eventivos y Grimshaw que solo aparecían en nominales eventivos).

Predicado Verbal

7. Predicado Verbal (fila 7 de la Tabla 5.1). Dado que la información en las 100.000 palabras de AnCora-Es analizadas no era suficiente para analizar este criterio, ha sido necesario ampliar la muestra de este tipo de predicados a las 500.000 palabras que contiene dicho corpus. Esta ampliación nos ha permitido analizar todas las ocurrencias (630 en total) de los predicados atributivos (‘ser’, ‘estar’ y ‘parecer’) y los predicados típicamente eventivos (‘tener lugar’, ‘ocurrir’, ‘comenzar’, ‘acabar’, ‘durar’, ‘llevar a cabo’) cuyos sujetos son alguna de las 817 nominalizaciones extraídas. A cada sustantivo deverbal que no estaba en la muestra inicial le hemos asignado una denotación para cada una de sus ocurrencias. La Tabla 5.1 muestra que los predicados atributivos tienden a elegir como sujetos SNs cuyos núcleos son sustantivos resultativos (75 %) mientras que los predicados típicamente eventivos no manifiestan una preferencia clara por ningún tipo de SN: el 44 % de ellos combina con nominalizaciones de resultado, el 41 % con nominalizaciones de evento y el 15 % con nominalizaciones subespecificadas. Estos resultados confirman parcialmente lo que mantienen los distintos autores: los sustantivos resultativos se combinan preferentemente con predicados atributivos.

En resumen, de este primer análisis lingüístico hemos concluido que la distinción semántica entre las nominalizaciones de evento y las de resultado no siempre es tan clara como parece en la bibliografía. Los criterios propuestos por los diferentes autores se adecuan bien a ejemplos contruidos *ad hoc* pero cuando se

aplican a muestras de lenguaje real no parecen funcionar tan bien: muchos de ellos no pueden aplicarse en todos los ejemplos y además, en muchas otras ocasiones encontramos criterios opuestos en un mismo ejemplo. Una vez dicho esto, es importante remarcar que estos criterios no son pruebas definitivas para distinguir las nominalizaciones eventivas y resultativas, sino indicadores que nos pueden ayudar a reforzar nuestra intuición semántica. De hecho, si hemos propuesto un tercer tipo denotativo subespecificado es porque existen casos en los que nuestra intuición semántica es insuficiente y el contexto de la frase también es insuficiente y, por lo tanto, los criterios para consolidar una de las dos denotaciones no son claros.

Respecto a los criterios evaluados, se confirman para el español como más concluyentes la clase verbal de la que deriva el sustantivo, la pluralización, el tipo de determinante, la preposición que introduce el complemento agentivo y la obligatoriedad del argumento interno (criterios del 1 al 5). Estos criterios se incorporaron como atributos en el léxico nominal, AnCora-Nom-v1. En cuanto a los dos restantes, es decir, la interpretación argumental de los SPs introducidos por la preposición 'por', adjetivos relacionales y determinantes posesivos (criterio 6), y los predicados verbales con los que se combinan los sustantivos (criterio 7), los resultados no son tan determinantes. Respecto al criterio 6, cabe destacar que los adjetivos relacionales parecen ser un buen indicador de la interpretación resultativa, mientras que los SPs introducidos por la preposición 'por' lo son de la interpretación eventiva (confirmando la hipótesis de Picallo (1999)). Sin embargo, el criterio no es suficientemente concluyente respecto a los determinantes posesivos, los resultados obtenidos no coinciden con ninguna de las hipótesis teóricas propuestas. En relación a los predicados verbales (criterio 7), la muestra analizada corrobora que los predicados atributivos tienden a combinarse con sustantivos resultativos pero no se confirma que los predicados típicamente eventivos prefieran la combinación con sustantivos eventivos.

Es interesante destacar que todos los criterios ofrecen rasgos morfosintácticos y semánticos particulares que refuerzan la identificación de sustantivos resultativos: si los sustantivos derivan de verbos de actividades y estativos, si el sustantivo aparece en plural, si el determinante que le precede es un indefinido, demostrativo o cuantificador, si la preposición que introduce a su complemento agentivo es 'de' o 'entre', si el argumento interno no se realiza ni se sobreentiende, si el argumento externo se realiza mediante adjetivos relacionales y si el predicado verbal con el que se combina el sustantivo es atributivo. Sin embargo, para la detección de la lectura eventiva solo uno de los criterios consolida esta lectura inequívocamente: cuando la preposición que introduce a su complemento agentivo es 'por' o 'por parte de'. Si además, tenemos en cuenta que el complemento agentivo es mayoritariamente opcional en la configuración de un SN, es muy difícil encontrar un criterio que dentro del SN permita reforzar la lectura eventiva.

Creemos que existen más rasgos para apoyar las nominalizaciones resultativas

porque son más próximas a los sustantivos no derivados y, como ellos, pueden admitir una amplia variedad de configuraciones: aparición en plural, diferentes tipos de determinantes, la posibilidad de aparecer sin complementos, etc. Las nominalizaciones eventivas, por lo contrario, dado que no son sustantivos prototípicos porque denotan acciones, al igual que los verbos, no admiten esta variedad de configuraciones: raramente aparecen sin complementos, admiten menos tipos de determinantes y aparecen en plural con mucha menos frecuencia. La mayoría de las configuraciones que las nominalizaciones eventivas admiten son también admitidas por las resultativas, pero no a la inversa, lo que explica que haya más criterios para la consolidación e identificación de la lectura resultativa que la eventiva.

En cuanto al resto de criterios -la derivación de verbos transitivos o inacusativos, la aparición en singular del sustantivo, la coocurrencia con el determinante definido, posesivo o sin determinante, la aparición del argumento interno y la combinación con predicados típicamente eventivos- no son rasgos determinantes para la identificación de una u otra lectura denotativa. Esto ocasiona que la clasificación de las nominalizaciones en eventivas resulte más difícil porque no se tienen criterios morfosintácticos determinantes que apoyen la semántica de la decisión, como se ha mencionado. Como los significados semánticos no siempre son claros, se dan numerosos casos en los que es imposible asignar una denotación concreta, de ahí la necesidad de un tercer tipo denotativo subespecificado. Este tipo denotativo se postuló ante la necesidad de marcar de alguna manera que el contexto oracional no era suficiente para establecer la denotación de la nominalización. Evidentemente, si ampliáramos el contexto de la oración al discurso, posiblemente se obtendrían nuevas características que nos permitirían desambiguar entre las dos lecturas. A parte de esto, la denotación subespecificada también pretende abarcar los casos en los que el tipo denotativo de la nominalización es ambigua, es decir, puede denotar ambas lecturas, la eventiva y la resultativa (4).

(4) [La **inversión** en investigación básica]_{SN} es el camino para el crecimiento.

5.1.2. Nuevos indicadores de la denotación

El análisis de las 3.077 ocurrencias de nominalizaciones centrado en la distinción semántica entre nominalizaciones de evento, resultado y de tipo subespecificado, nos permitió encontrar pistas nuevas que nos ayudaban a consolidar una de estas lecturas. De hecho, la mayoría de los nuevos indicadores sirven para reforzar la lectura eventiva que, como hemos visto, tenía menos criterios de la bibliografía que la apoyaran.

Una de las pruebas que más nos sirvió para detectar las nominalizaciones eventivas es la posibilidad de parafrasearlas por una estructura clausal. Recuérdese los

ejemplos del Capítulo 1 (9), (10) (11) y (12), repetidos aquí como (5), (6) (7) y (8).

- (5) **Se ha ampliado** el capital de la empresa en un 20 %.
- (6) [**La ampliación**<evento> del capital de la empresa en un 20 %] SN.
- (7) Se han vendido [**muchas traducciones**<resultado> de su último libro]SN.
- (8) Se han vendido [**muchos libros traducidos** de su último título] SN.

Si un SN (6) permite la paráfrasis por una oración (5), se considera que es un SN cuyo núcleo es una nominalización eventiva. Esta paráfrasis oracional, sin embargo, resulta imposible si la nominalización tiene una interpretación resultativa como en (7) -‘traducciones’ se refiere al objeto concreto, es decir, al libro traducido. En este sentido, las nominalizaciones deverbales resultativas solo pueden ser paráfrasis de otros SNs que denoten objetos (8).

Otra prueba importante que nos ayuda establecer la denotación es si la nominalización admite un complemento agentivo introducido por las preposiciones ‘por’ o ‘por parte de’. Hacemos uso de este criterio porque es el más informativo para consolidar la lectura eventiva de las nominalizaciones eventivas pero también es muy opcional y está muy poco representado en el corpus. De tal modo, que ante un SN sin este tipo de complemento agentivo (9), se trata de insertarlo y comprobar que funciona (10).

Criterio Agente

- (9) Se ha informado de [la **rebaja** de los sueldos de los funcionarios]SN.
- (10) Se ha informado de [la **rebaja** de los sueldos de los funcionarios por parte del Gobierno]SN.

Los anotadores se sirvieron de estas dos pruebas para establecer la denotación de las nominalizaciones, además de su criterio semántico.

También encontramos otro tipo de indicadores que nos ayudan a establecer una denotación, son los llamados selectores, elementos que ayudan a seleccionar la denotación de la nominalización. Los selectores pueden ser de dos tipos: (i) los selectores externos, es decir, los elementos que desde fuera del SN indican la denotación de la nominalización; y (ii) los selectores internos, es decir, prefijos de la nominalización que indican un tipo concreto de denotación. Como selectores externos incluimos preposiciones (11), sustantivos (12), adjetivos (13), verbos (14) y adverbios (15).

Selectores

- (11) Durante [la **presentación**<evento> del libro]SN, él abogó por la formación de los investigadores en innovación tecnológica.
- (12) El gobierno checo quiere comenzar el proceso de [**privatización**<evento> de este banco]SN.

- (13) Una de las primeras formas de piel tuvo que ser algo así como una membrana, resultante d[el **endurecimiento**<evento> de la sustancia celular]SN.
- (14) [La **discusión**<evento>]SN empezó en seguida, porque olvidaron cerrar la puerta.
- (15) Una generación en vías de [**extinción**<evento>]SN.

La preposición ‘durante’ con su marcado valor durativo nos da la pista en (11) para considerar a la nominalización ‘presentación’ como evento. En (12) observamos que algunos sustantivos como por ejemplo el nombre ‘proceso’ induce a la lectura eventiva de ‘privatización’. Del mismo modo, adjetivos como ‘resultante’ en (13) influyen en la lectura eventiva de la nominalización ‘endurecimiento’. También los verbos son selectores muy potentes; por ejemplo, si el sujeto o complemento directo de un verbo como ‘empezar’ contiene una nominalización, esa nominalización tenderá a ser eventiva (14). Finalmente, una locución adverbial como ‘en vías de’ en (15) apunta a la lectura eventiva de ‘extinción’.

Además de los selectores externos, también encontramos características morfológicas de las nominalizaciones (selectores internos) que también pueden influir en la denotación de la nominalización. Por ejemplo, una nominalización con el prefijo ‘re-’ con un significado reiterativo suele ser eventiva (16) puesto que el significado reiterativo solo se puede aplicar a bases que denotan acciones.

- (16) Hoy [la **reubicación**<evento> del ex ministro]SN no resulta fácil.

El conjunto de nuevos indicadores nos ayuda a establecer una clasificación semántica de las nominalizaciones según su denotación, independientemente de los criterios de la bibliografía que son los que se evalúan. El único inconveniente de estos criterios (las dos pruebas semánticas y los selectores) es que no pueden representarse como atributos en el léxico AnCora-Nom-v1, por lo que más tarde no se implementarán como rasgos del Clasificador ADN.

5.2. Denotación: análisis computacional

A partir del estudio lingüístico realizado, se elaboró manualmente un léxico, AnCora-Nom-v1, en el que se incluían las 817 entradas correspondientes a los lemas estudiados. Cada entrada se organizó en diferentes sentidos (un total de 1.121) que fueron establecidos en función de las diferentes denotaciones asociadas. Además del tipo denotativo, cada uno de los sentidos nominales contenía los siguientes atributos: el lema y la clase verbal del verbo del que deriva la nominalización; los constituyentes del SN cuyo núcleo es la nominalización, especificando

si son argumentales o no y qué clase de argumentos son; el tipo de determinante que aparece en los ejemplos asociados a aquel sentido de la nominalización; y si el sustantivo en aquel sentido determinado aparece en plural. También se asociaba a cada sentido los *synsets* correspondientes a la versión 1.6 del WordNet español, se señala si forman parte de una construcción lexicalizada y se especifica el tipo de nominalización (en este caso, son todas deverbales). Además, cada sentido tiene asociadas las oraciones del corpus que ejemplifican los atributos anotados (un total de 3.077 ejemplos)². Este léxico es una versión inicial y parcial del léxico AnCora-Nom que contiene todas las nominalizaciones del corpus AnCora-Es (1.655 en total), que es uno de los recursos finales del proceso de investigación que aquí se presenta. Se debe tener en cuenta que tanto el corpus AnCora-Es como el léxico AnCora-Nom han sido utilizados reiterativamente en los distintos procesos y que han sido completados en diferentes etapas hasta llegar a las versiones finales que presentamos en los Capítulos 8 y 9 respectivamente. Por lo tanto, en esta subsección al mencionar AnCora-Nom, nos referiremos a la versión primera de este léxico a la que nos referiremos como AnCora-Nom-v1.

A partir de AnCora-Nom-v1, se realizaron una serie de experimentos cuyo objetivo era doble: por una parte, disponer de un marco para refrendar empíricamente las hipótesis lingüísticas y evaluar cuantitativamente la importancia de los diferentes criterios que consideramos pertinentes para el español (tanto individualmente como combinados); y por otra parte, sentar las bases para la construcción de un sistema automático que clasifique un nombre susceptible de constituir una nominalización deverbal como evento o resultado en función del contexto de aparición (Peris et al., 2009). Se utilizaron técnicas de ML para llevar a cabo tanto el análisis de los rasgos como la construcción del clasificador. Nuestra hipótesis es que la combinación de los criterios establecidos en el Capítulo 2 y analizados en la sección anterior debiera contribuir a aumentar la precisión en la tarea de clasificación. Como herramienta de aprendizaje se utilizó el conocido paquete Weka (Witten and Frank, 2005). El tipo de aprendizaje fue supervisado ya que disponíamos del corpus de entrenamiento etiquetado manualmente (las 3.077 ocurrencias anotadas). La evaluación se llevó a cabo utilizando validación cruzada a partir de 10 particiones aleatorias (*10 fold Cross-validation*)³. De entre los clasificadores que

Experimentos sobre
AnCora-Nom-v1

²Si una oración del corpus contenía más de una nominalización, dicha oración se repite como ejemplo en cada una de las entradas léxicas correspondientes.

³En este método de evaluación, la muestra de datos se divide aleatoriamente en N submuestras. De estas N submuestras, solo una se conserva como muestra de datos para la evaluación del modelo y el resto ($N-1$) se usa como muestra de datos de entrenamiento. Este proceso es repetido N veces y en cada una de ellas se usa una de las N submuestras como muestra de datos para la evaluación del modelo. De estos N resultados se obtiene una media, que es la evaluación del modelo (McLachlan et al., 2004). En nuestro caso tomamos $N=10$. El método es especialmente útil cuando se dispone de una muestra pequeña ya que toda ella se utiliza para aprender en alguna de las N iteraciones.

Weka ofrece se seleccionó J48.Part, la versión en reglas del clasificador de árboles de decisión C4.5 (Quinlan, 1993). Dicha elección está fundamentada por dos motivos: i) un análisis inicial con otros clasificadores más potentes (o al menos más robustos) como los SVM o el *Adaboost* no pareció dar resultados significativamente mejores; y ii) el modelo de clasificación aprendido consiste en una secuencia de reglas simbólicas cuya interpretación por el lingüista es posible. De hecho, la interpretación de estas reglas simbólicas nos ha permitido la detección de nuevos indicadores para establecer la distinción entre evento y resultado en las nominalizaciones deverbales. A continuación nos centramos en los experimentos llevados a cabo para refrendar empíricamente las hipótesis establecidas (Subsección 5.2.1) y para terminar detallaremos los nuevos indicadores para establecer la distinción entre evento y resultado en las nominalizaciones deverbales obtenidos a partir de la observación de las reglas simbólicas (Subsección 5.2.2).

5.2.1. Experimentos para la evaluación de AnCora-Nom-v1

ADN-Classifier-v1

En estos experimentos se utilizaron como rasgos las propiedades contenidas en las entradas léxicas de AnCora-Nom-v1 y dado que las entradas se organizan en sentidos, los ejemplos de aprendizaje corresponden a sus 1.121 sentidos. Por lo tanto, los experimentos se realizaron a nivel de sentido. La Figura 5.1 nos muestra el proceso de realización de estos experimentos. A partir de AnCora-Nom-v1, se extraen los atributos a nivel de sentido, que incluyen el resultado (la supervisión), a partir de los cuales aprende Weka, dando lugar a un modelo de clasificación que posteriormente Weka utiliza en modo clasificación para asignar a los distintos ejemplos a clasificar en una de las tres denotaciones establecidas (evento, resultado, subespecificado) o en una lexía. Este modelo de clasificación constituye el primer estadio del clasificador ADN, lo que conocemos por ADN-Classifier-v1. Además, dado que en AnCora-Nom-v1 están básicamente codificados los criterios establecidos como pertinentes en la sección anterior, el resultado de la clasificación sirve también para evaluar dichos criterios (Peris et al., 2009).

Rasgos léxicos de aprendizaje

En la Tabla 5.2 se recogen los rasgos utilizados en el aprendizaje. En la columna 1 se indica la clase de rasgo: la clase verbal de la que deriva la nominalización, la posibilidad de aparecer en plural, el tipo de determinante, el tipo de nominalización deverbal, si forma parte de algún tipo de lexía y los diferentes constituyentes que aparecen en el SN cuyo núcleo es la nominalización deverbal. La columna 2 indica el número de valores del rango (conjunto de valores posibles) de cada uno de los rasgos. En algunos casos el valor de un rasgo está indefinido, por ello se ha añadido el valor “nil” a cada uno de los rangos. Debido a la excesiva dispersión de los valores posibles en algunos casos, que conduce a la insuficiente representatividad (*data sparseness*) de los mismos y, por lo tanto a una degradación en el proceso de aprendizaje, se ha añadido la posibilidad de agrupar algunos de estos

Agrupación de rasgos

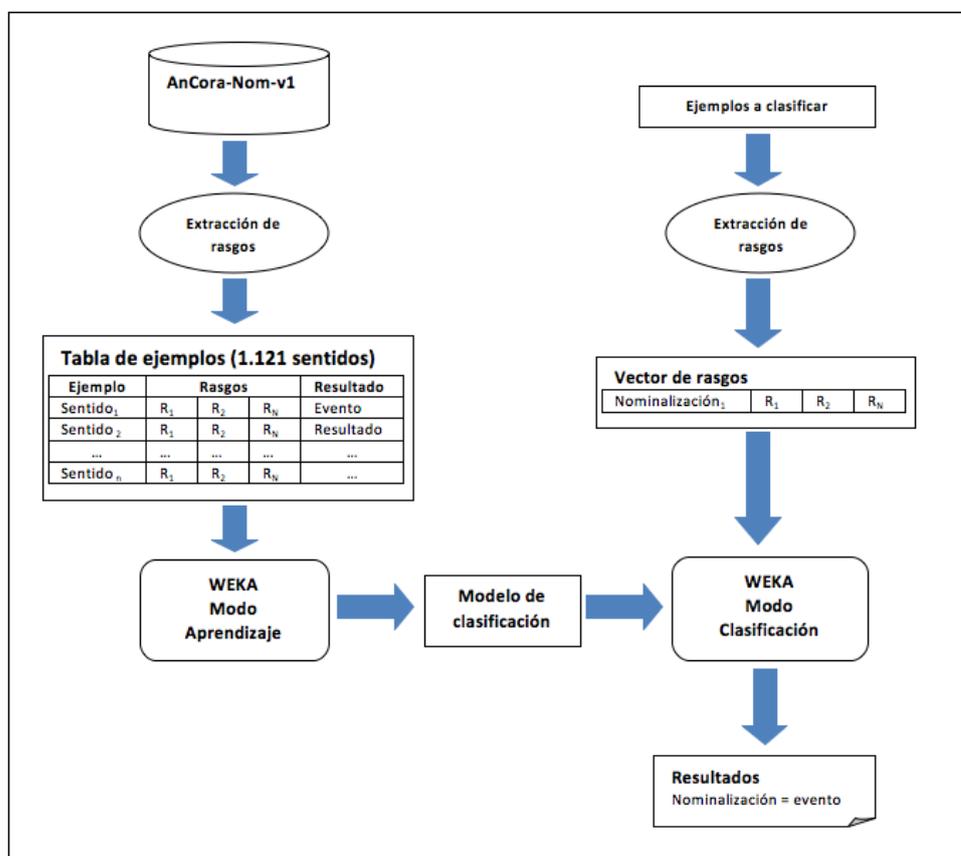


Figura 5.1: Esquema de los experimentos computacionales para la verificación de los criterios

valores para facilitar el aprendizaje. La columna 3 presenta el tamaño del rango para los valores agrupados. El caso más interesante de esta agrupación es el del rasgo SP: en los SPs existen 101 valores posibles resultantes de la combinación de las diferentes posiciones de los argumentos (arg0, arg1, arg2, arg3, arg4, argM), las diferentes preposiciones ('de', 'por', 'entre', 'con', 'para', etc.) y los papeles temáticos (agente, paciente, tema, etc.) y este número de valores posibles es demasiado elevado para los 1.121 ejemplos de aprendizaje disponibles. En este caso, se han considerado dos agrupaciones: una a nivel de número de argumento (arg0, arg1, arg2, arg3, arg4, argM, además del valor no argumental, RefMod, que proporciona, pues, 7 valores posibles) y otra más fina que agrupa la información argumental y la preposición involucrada (arg0-con, arg0-de, etc. dando lugar a 60 valores posibles). Para cada uno de los rasgos se ha realizado también una descomposición binarizada, es decir, se ha añadido para cada valor posible del

Binarización de rasgos

rango un rasgo binario que indicara cuando el valor correspondía a dicho rasgo⁴. Esta técnica permite también hacer frente al problema de la dispersión de datos descrito anteriormente. En general, la inclusión de rasgos binarizados ha resultado beneficiosa tal como indican los resultados de los experimentos en la Tabla 5.3. La columna 4, finalmente, incluye ejemplos de pares atributo-valor para cada rasgo.

Rasgos	Rango	Rango agrupado	Ejemplos
Clase Verbal	14	12	els = b2
Plural	2	-	plural = yes
Determinantes	74	15	espec = def
Tipo	4	-	tipo = nombre
Lexías	6	-	lexía = centro de acogida
SP1	101	7	SP = arg1-de-tem / arg1
SP2	101	60	SP = arg2-con-ins / arg2-con
SA	9	5	SA = arg0-agt
SN	2	-	SN = argM-loc
SADV	2	-	SAdv = argM-tmp
O.Sub	1	-	O.Sub = RefMod
Poss	5	2	Poss = arg1-pat
GRel	4	3	Rel = arg1-tem

Tabla 5.2: Rasgos utilizados en los experimentos a nivel de sentido para la validación empírica de AnCora-Nom-v1

Resultados

La Tabla 5.3 recoge los resultados obtenidos. Para llevar a cabo la evaluación se ha confeccionado un caso base (*baseline*) que se limita a devolver la clase más frecuente, esto es, la clase resultativa. El caso *simples* utiliza los rasgos de la Tabla 5.2 en su versión escalar, sin binarizar ni agrupar; 2) el caso *binarized* usa los mismos rasgos añadiendo ahora los correspondientes binarizados (en general, se ha adoptado el criterio de no eliminar los anteriores al refinar los rasgos de forma que los casos suficientemente representados puedan ser usados por el mecanismo de aprendizaje y los rasgos correspondientes incorporados al clasificador); y 3)

⁴Por ejemplo, el rasgo *lexía* admite seis valores posibles (“nominal”, “verbal”, “adjetival”, “preposicional”, “adverbial” y “conjuntiva”), su expresión binarizada consiste en seis rasgos (*lex-nom*, *lex-verb*, *lex-adj*, *lex-adv*, *lex-prep* y *lex-conj*) con dos valores posibles, TRUE, FALSE.

5. LA DENOTACIÓN EN LAS NOMINALIZACIONES DEVERBALES: ESTUDIO
EMPÍRICO

los siguientes casos van incorporando rasgos agrupados de forma incremental. En la segunda columna se contabiliza el número de rasgos utilizado en cada caso. La tercera columna informa del número de reglas aprendidas y usadas por el clasificador en cada caso. La cuarta columna presenta la corrección (*accuracy*), es decir, el número de ejemplos bien clasificados respecto al número total de ejemplos. Finalmente, la quinta columna informa del decrecimiento del error respecto al caso base (*baseline*).

Rasgos	Nº de Rasgos	Nº de Reglas	Corrección	Δ error
Caso Base	-	1	71,98 %	
<i>Simples</i>	12	24	82,07 %	10,09 %
<i>Binarizados</i>	12	32	83,22 %	11,24 %
Tipo	19	27	83,40 %	11,42 %
Clase verbal	34	40	83,03 %	11,05 %
Determinante	214	40	84,56 %	12,58 %
SP 1	134	30	84,03 %	12,05 %
SP 2	211	40	83,76 %	11,78 %
SA	221	40	84,47 %	12,49 %
Poss	231	38	84,48 %	12,50 %
GRel	247	30	84,57 %	12,59 %

Tabla 5.3: Resultados de los experimentos a nivel de sentido para la validación empírica de AnCora-Nom-v1

Los resultados obtenidos en estos experimentos son positivos: se clasifican correctamente el 84,57 % de los sentidos nominales, es decir, existe un 12,59 % de mejora respecto al caso base (*baseline*), es decir, una disminución de la tasa de error de casi un 50 % (15,43/28,02). A su vez, este resultado corrobora que los datos anotados en AnCora-Nom-v1 permiten detectar la distinción entre la lectura eventiva o resultativa de las nominalizaciones deverbales.

En concreto, se observó que la utilización de los rasgos aunque sea a nivel simple produce un incremento notable de la precisión del clasificador de denotaciones (del 71,98 % al 82,07 %), lo que supone una validación empírica de los datos anotados en AnCora-Nom-v1 y, por lo tanto, se refrendan los rasgos utilizados para establecer la distinción entre evento y resultado. También la binarización de los rasgos supone una mejora significativa (hasta el 83,22 %). La inclusión de

un número creciente de rasgos agrupados es siempre positiva aunque no todas las agrupaciones contribuyen igualmente y no siempre su combinación supone una mejora. Además, las diferencias entre ellas no son estadísticamente significativas en todos los casos.

Análisis de errores

Se llevó a cabo un análisis de los errores para estos experimentos con el objetivo de detectar dónde cabía una mejora. En la Tabla 5.4 se presentan los resultados obtenidos para cada clase denotativa en cuanto a la precisión, la cobertura y la F1⁵. En la Tabla 5.5 se muestra la matriz de confusión⁶.

De estos resultados, cabe destacar que el sistema clasifica mucho mejor los sustantivos resultativos (92,7 % de F1) que los eventivos (62,7 % F1) y subespecificados (34,5 % F1). Esto se debe a que existen más rasgos que permiten identificar la clase de resultativos (pluralización, tipo de determinante, clase verbal, adjetivos relacionales). En cambio, en el caso de los subespecificados, como no se dispone de ningún rasgo particular que los identifique (de ahí su clasificación como subespecificados), el sistema no consigue una clasificación óptima. Entre los clasificados como resultativos, el 24,3 % corresponde a errores de la clasificación manual en el léxico (es decir, que se clasificaron como subespecificados pero en el análisis de errores se comprobó que eran resultativos), por lo que podríamos considerar que este porcentaje en realidad está bien clasificado automáticamente. El 40,5 % de los casos se explican porque se trata de sentidos subespecificados que, o bien no tienen complementos asociados en la entrada, o bien estos complementos no son argumentales (s.a = RefMod, sp = RefMod, OSub= RefMod), y esta casuística tiende a aparecer mayoritariamente en sentidos resultativos, de ahí la confusión en la clasificación. En cuanto al 35,2 % restante, son casos cuyos atributos no representan mayoritariamente la clase de subespecificados, sino que se trata de rasgos que coinciden con la clase de resultativos. De ahí que se clasifiquen como resultativos cuando son subespecificados. Este mismo problema (la coincidencia de rasgos que pueden caracterizar ambas clases) explica los casos de resultativos clasificados incorrectamente como subespecificados. Este mismo argumento, que dos clases compartan la misma casuística de atributos, es válido tanto para los casos subespecificados clasificados como eventivos, como para los casos eventivos clasificados como subespecificados. En el caso de los eventivos, el índice de acierto es menor que en el caso de los resultativos porque también

⁵La precisión (*precision*) y la cobertura (*recall*) son medidas complementarias. La F1 pretende ser una medida global de la calidad del clasificador. La F1 es la media armónica ponderada de las dos medidas básicas. En nuestro caso, el peso de cada medida básica es el mismo (0,5) de forma que damos la misma importancia a precisión y cobertura.

⁶La matriz de confusión es una matriz de dos dimensiones cuyas columnas corresponden a la clasificación producida por el sistema automático y las filas a los valores correctos. Por ejemplo, de los 807 ejemplos de tipo resultado, 765 han sido correctamente clasificados, a 20 de ellos se les ha asignado la etiqueta eventiva y a 22 la de subespecificado.

5. LA DENOTACIÓN EN LAS NOMINALIZACIONES DEVERBALES: ESTUDIO EMPÍRICO

es menor el número de rasgos identificativos de esta clase de nombres (por-SP, posesivo argumental). En concreto, los 23 casos eventivos erróneamente clasificados como resultativos aparecen con un SP que es arg1 y con complementos no argumentales, característica compartida mayoritariamente por la clase de resultativos, y de ahí su incorrecta clasificación. Finalmente, los 20 casos de resultativos clasificados como eventivos aparecen con un único SP que es arg1, mayoritariamente representativo de la clase de eventivos. La F1 más alta (99,3 %) lo presenta la clase de las lexías no-nominales marcadas explícitamente en AnCora-Nom-v1, de ahí el alto porcentaje de acierto.

Clase	Precisión	Cobertura	F1
R	0,906	0,948	0,927
SE	0,515	0,260	0,345
E	0,563	0,708	0,627
L	1	0,986	0,993
Global	0,82	0,84	0,83

Tabla 5.4: Análisis de errores de los experimentos a nivel de sentido para la validación empírica de AnCora-Nom-v1. Leyenda: R= resultado, SE = subespecificado, E=evento y L = lexía.

Clasificación correcta \Rightarrow	Clasificación del sistema \Downarrow				Total valores correctos
	R	SE	E	L	
R	765	22	20	0	807
SE	55	34	42	0	131
E	23	10	80	0	113
L	1	0	0	69	70
Total Sistema	844	66	142	69	1.121

Tabla 5.5: Matriz de confusión de los experimentos a nivel de sentido para la validación empírica de AnCora-Nom-v1. Leyenda: R= resultado, SE = subespecificado, E=evento y L = lexía.

El análisis de errores mostró que eran necesarios criterios adicionales para la distinción de evento y resultado en las nominalizaciones deverbales, sobre todo criterios que sirvieran para detectar la lectura eventiva, donde la dificultad es

mayor para el clasificador. Por este motivo, decidimos enriquecer el modelo obtenido con rasgos extraídos del corpus AnCora-Es, como son el tiempo y la clase semántica del verbo principal de la oración donde se encuentra la nominalización⁷, la función sintáctica del SN en el que se encuentra la nominalización, si la nominalización constituye o no una entidad con nombre, y algunas combinaciones de estos rasgos (Peris et al., 2010a).

Rasgos	Rango	Ejemplos
Tiempo Verbal	5	tense=past
Clase Verbal	14	els = a2
Función sintáctica	9	func = suj
Entidad con Nombre	2	ne = true
Tiempo Verbal + Función Sintáctica	45	tense=past+func=cd
Clase Verbal + Función Sintáctica	126	els=a1+func=cd

Tabla 5.6: Rasgos contextuales empleados en los experimentos a nivel de corpus.

Se tiene que de tener en cuenta que en este caso los ejemplos para el aprendizaje son los 3.077 ejemplos del corpus asociados a los 1.121 sentidos, es decir, los ejemplos de aprendizaje no son ya los sentidos sino cada una de sus ocurrencias en AnCora-Es. La extracción de los rasgos a nivel de corpus se llevó a cabo a partir de la información representada en los árboles sintácticos de esos 3.077 ejemplos del corpus AnCora-Es. Para esta tarea se utilizó la herramienta Tgrep2⁸ que permite la manipulación e inspección de árboles de análisis en formato *treebank* de forma simple y eficiente. Se han implementado 108 reglas⁹ que nos han permitido extraer de los árboles de análisis de AnCora-Es la información relativa a los rasgos contextuales presentados en la Tabla 5.6.

A continuación, se incluye un ejemplo para ilustrar el proceso. Un rasgo que hemos considerado interesante incluir es la aparición de la nominalización deverbal en posición de sujeto. El patrón de Tgrep2 que nos extrae esta información

⁷En la Tabla 5.6 cuando hablamos de tiempo y clase verbal, nos referimos al del verbo principal de la oración donde aparece la nominalización.

⁸<http://tedlab.mit.edu/~dr/TGrep2/>

⁹El conjunto de reglas Tgrep implementadas también se puede consultar en el siguiente enlace: <http://clic.ub.edu/corpus/en/documentation>

para el nombre ‘construcción’ es:

“sn < (“grup.nom” <(n </construcción/)) </func *subj*/”

El patrón genérico aplicable a cualquier nominalización sería:

“sn < (“grup.nom” <(n </arg***/)) </func *subj*/”

Este patrón se puede parafrasear como ‘búsqueda de un SN con la función de sujeto que domine inmediatamente un grupo nominal que a su vez domine el nombre que contiene el sustantivo deverbal, en este caso, ‘construcción’¹⁰.

En la Tabla 5.7 se presentan los resultados de la incorporación de los atributos del corpus extraídos al modelo desarrollado. Estos son aparentemente muy superiores pero en realidad la mejora es pequeña. El caso base en este caso es del 82 % frente al 72 % anterior. Esto se debe a que cuando se pasa del marco que toma como unidad el sentido al marco cuya unidad es la oración, el número de ejemplos para el aprendizaje se incrementa de 1.121 a 3.077 y, además, la proporción real (la que existe en el *treebank*) de ocurrencias de nombres resultativos también aumenta. Por ello, el 93,56 % de corrección atribuido a la clase “solo lemas”, correspondiente a la última columna de la Tabla 5.7, debe considerarse en relación al caso base (columna 5 de la Tabla 5.7). Siendo los sentidos del léxico la unidad de clasificación, la mejora sobre el caso base era de un 12,59 % y ahora es de un 11,56 %, es decir, no hay mejora. En la tercera fila, se introducen los rasgos correspondientes a la aparición del nombre en función de sujeto y complemento directo, en este caso la mejora no es estadísticamente significativa. Tampoco lo es la introducción (fila 4) del resto de funciones sintácticas, que además suponen un descenso en el nivel de corrección. Este fenómeno se repite también con la introducción como rasgo del tiempo verbal (fila 5) y de la entidad con nombre (fila 6).

En resumen, los resultados obtenidos no fueron muy esperanzadores puesto que añadiendo la información de los rasgos del corpus, la corrección bajaba en un punto respecto a los resultados obtenidos con rasgos obtenidos exclusivamente del léxico. Esto nos demostró que existe información en el léxico que parece crucial para establecer la distinción entre evento y resultado, como es la información sobre la estructura argumental y el verbo base de la nominalización. De ahí, que para anotar la denotación de todas las nominalizaciones del corpus AnCora-Es con un clasificador automático, necesitamos para construirlo información sobre la estructura argumental. A partir de estas conclusiones, se consideró necesario anotar la

¹⁰En el proceso de búsqueda el parámetro *arg**** se sustituye por cada una de las nominalizaciones.

Rasgos	Nº de Rasgos	Nº de Reglas	Corrección	Δ error
Caso Base	-	1	82 %	
Solo lemas	251	61	93,56 %	11,56 %
SUJ y CD	258	59	93,63 %	11,63 %
Otras funciones sintácticas	258	52	93,40 %	11,40 %
Tiempo verbal	258	50	93,30 %	11,20 %
Entidad con Nombre	258	48	92,80 %	10,80 %

Tabla 5.7: Resultados de los experimentos a nivel de sentido añadiendo rasgos de AnCora-Es a los rasgos de AnCora-Nom-v1

estructura argumental de todas las nominalizaciones del corpus AnCora-Es, tal y como se ha explicado en el capítulo anterior. Además, parece claro que para desarrollar un clasificador de nominalizaciones según su denotación que sea capaz de prescindir de la información del léxico, necesitamos incrementar también la muestra de datos para el aprendizaje. Por este motivo, necesitamos anotar la denotación de las nominalizaciones deverbales del corpus AnCora-Es. Por lo tanto, era necesario anotar la denotación en todas las ocurrencias de nominalizaciones deverbales de AnCora-Es al completo. Dado que esto implica un notable aumento de las ocurrencias a anotar (23.431 ocurrencias en comparación con las 3.077 hasta ahora anotadas), necesitábamos automatizar el proceso. Para tal propósito se adaptó el modelo de clasificación a nivel de sentido aprendido (ADN-Classifier-v1) a un modelo de clasificación a nivel de lemas (ADN-Classifier-v2), con el objetivo de que se pudieran clasificar automáticamente según su tipo denotativo las ocurrencias de nominalizaciones del corpus AnCora-Es, cuyos sentidos eran desconocidos. Este nuevo modelo parte de los siguientes recursos: 1) el léxico AnCora-Verb para obtener los rasgos relacionados con los verbos correspondientes a las nominalizaciones; 2) el corpus AnCora-Es-v2 al completo, del que se obtienen distintos rasgos morfosintácticos y semánticos; y 3) el recién creado léxico AnCora-Nom-v2 (véase la Sección 1.3 del Capítulo 1 y la Sección 9.1 del Capítulo 9), del que se obtiene, entre otras, la información sobre la estructura argumental de todos los lemas de las nominalizaciones del corpus. Sin embargo, dado que trabajamos a nivel de lema, prescindimos de toda información que era específica de un sentido determinado y solo tuvimos en cuenta la información compartida por todos los sentidos de una mismo lema, lógicamente esta granularidad más grosera tuvo el coste de una caída en la precisión de ADN. El nuevo modelo de clasificación a

nivel de lema (ADN-Classifier-v2) se utilizó para la anotación del tipo denotativo en el corpus AnCora-Es. Con el objetivo de evaluar el rendimiento de este modelo, que consiguió 85,27 % de corrección sobre un caso base de 83,92 % (según se indica en la Tabla 7.1 de la Sección 7.3), se validó manualmente el corpus (Peris et al., 2010b), dando lugar a una nueva y definitiva versión de AnCora-Es (-v3) (véase el Capítulo 8). A partir de este corpus manualmente validado se generó la versión final del léxico AnCora-Nom (-v3) (véase el Capítulo 9) que incluye también información sobre el tipo denotativo para todas las entradas léxicas. A partir de estos recursos se creó la versión final del Clasificador ADN (-v3) (véase el Capítulo 6).

5.2.2. Criterios obtenidos a partir de la observación de las reglas del modelo de clasificación

La interpretación de la reglas simbólicas desarrolladas por el clasificador nos ha permitido la detección de nuevos indicadores para establecer la distinción entre evento y resultado en las nominalizaciones deverbales. Concretamente, los nuevos criterios surgen de la observación de reglas en las que interaccionan tres tipos de información básica: la clase verbal del verbo correspondiente a la nominalización, los argumentos realizados y los constituyentes que realizan dichos argumentos. Recuérdese que los verbos transitivos, que se correspondían con las realizaciones del léxico verbal AnCora-Verb, y los inacusativos, que se correspondían con los logros del léxico verbal AnCora-Verb, son los que pueden dar lugar a los tres tipos de denotación (evento, resultado, subespecificado). Los estados y las actividades (verbos estativos e inergativos respectivamente en AnCora-Verb) solo parecen dar lugar a nominalizaciones resultativas, lo que las reglas también confirman. Por este motivo, nos centramos en las nominalizaciones derivadas de verbos que pertenecen a las clases semánticas de las realizaciones o logros.

Realizaciones (Clase A). Los sentidos verbales que pertenecen a esta clase semántica pueden dar lugar a nominalizaciones resultativas, eventivas y subespecificadas. La lectura de la nominalización depende de qué argumentos se realizan en el SN y cuáles son los constituyentes que los explicitan. Por ejemplo, una nominalización tiende a ser un evento cuando tiene un arg1-pat realizado por un determinante posesivo (17).

- (17) [[Su]_{Poss-arg1-pat} **persecución**_{<evento>}]_{SN} es uno de esos atractivos marginales que aún le quedan a la Vuelta [...].

La Tabla 5.9 determina el tipo denotativo según los argumentos realizados y los constituyentes empleados. Por ejemplo, en la segunda fila se observa que si no

se ha realizado ningún argumento la nominalización se considera resultativa (18), sin embargo en la tercera fila se muestra que si el arg2 se realiza mediante un SP o un Grel (pronombre relativo genitivo) la nominalización es subespecificada, si además el arg1 se realiza por los mismos constituyentes (fila 4) la nominalización se considera eventiva (20).

- (18) Es heroico levantarse todas las mañanas a luchar contra [la **desesperación**<resultado>]SN.
- (19) [La **conversión**<subespecificado> [a euros]SP-arg2-]SN no es efectiva.
- (20) [La **apertura**<evento> [de la red de telefonía local de Telefónica]SP-arg1-pat [a otros operadores]SP-arg2-]SN.

Logros (Clase B). Los sentidos verbales pertenecientes a la clase de los logros se realizan en estructuras sintácticas inacusativas (estructuras sintácticas sin arg0), por lo que la denotación de la nominalización que derivan depende de la realización sintáctica de los arg1 y arg2 (Véase la Tabla 5.8). Por ejemplo (fila 5 de la Tabla), si tanto el arg1 como el arg2 se realizan la nominalización se considera eventiva (21), mientras que si solo se realiza uno de ellos, en el ejemplo (22) el arg1, la nominalización se considera subespecificada.

- (21) Los corredores de la Vuelta Ciclista a Alemania darán los últimos golpes de pedal de [[su]Poss-arg1-pat **entrada** <evento> [a la capital]SP-arg2-loc]SN.
- (22) Para las elecciones, se espera [la **llegada**<subespecificado> [de más de 450 observadores extranjeros]SP-arg1-tem]SN.

arg1	arg2	Denotación
No realizado	No realizado	Resultado
No realizado	Realizado	Subespecificado
Realizado	No Realizado	Subespecificado
Realizado	Realizado	Evento

Tabla 5.8: Tipo denotativo según la realización argumental de nominalizaciones derivadas de verbos de la clase semántica de los logros

5.3. Conclusiones

En este capítulo se ha presentado el estudio empírico sobre la denotación de las nominalizaciones deverbales del español. Este estudio nos ha permitido, por una parte, evaluar qué criterios de la bibliografía son válidos para el español y por otra parte, nos ha permitido inferir nuevas pruebas que nos ayudan a distinguir entre los distintos tipos denotativos posibles. Estas nuevas pruebas son tanto el resultado del análisis lingüístico llevado a cabo (Sección 5.1.2) como de las observaciones de las reglas simbólicas generadas por el sistema automático (Sección 5.2.2). Además, las técnicas de ML utilizadas para el análisis computacional, han sentado las bases para el desarrollo de un sistema automático de clasificación de las nominalizaciones deverbales según su denotación, el clasificador ADN, que se presenta en el siguiente capítulo.

arg0	arg1	arg2	Denotación
No realizado	No realizado	No realizado	Resultado
No realizado	No realizado	Realizado: SP/Grel	Subespecificado
No realizado	Realizado: SP/Grel	Realizado: SP/Grel	Evento
Realizado: SP-por agente	Realizado: SP/Grel/Poss	No realizado o Realizado: SP/Grel/Poss	Evento
Realizado: SP-de agente	Realizado: SP/Grel/Poss	No realizado o Realizado: SP/Grel/Poss	Resultado
Realizado: SP-por/de agente	Realizado: SP/Grel/Poss	No realizado o Realizado: SP/Grel/Poss	Subespecificado
Realizado: SP-por/de agente	Realizado: SP/Grel/Poss	Realizado: SP/Grel/Poss	Evento
Realizado: SP-por/de causa	Realizado: SP/Grel/Poss	No realizado	Subespecificado
Realizado: SP-por/de causa	Realizado: SP/Grel/Poss	Realizado: SP/Grel/Poss	Subespecificado
Realizado: cualquier constituyente	Realizado: Poss paciente	Realizado o No realizado	Evento

Tabla 5.9: Tipo denotativo según la realización argumental de nominalizaciones derivadas de verbos de la clase semántica de las realizaciones

CAPÍTULO 6

CLASIFICADOR ADN: CLASIFICADOR AUTOMÁTICO DE NOMINALIZACIONES DEVERBALES SEGÚN SU DENOTACIÓN

En este capítulo presentamos la versión final del clasificador automático de las nominalizaciones deverbales del español según su denotación en español (ADN-Classifier-v3) (Peris et al., 2012). Este clasificador tiene como punto de partida los experimentos computacionales realizados para el estudio empírico de las nominalizaciones deverbales (ADN-Classifier-v1), después se desarrolló con el objetivo de anotar automáticamente la denotación de las nominalizaciones deverbales del corpus AnCora-Es (ADN-Classifier-v2) y, finalmente, ha dado como resultado un clasificador capaz de trabajar en diferentes escenarios, es decir, en diferentes condiciones de clasificación (ADN-Classifier-v3). En otras palabras, la construcción del clasificador ADN ha sido un proceso incremental del que presentamos el resultado final: ADN-Classifier-v3. Tras la anotación automática de la estructura argumental (Véase el Capítulo 4) y la denotación (Véase la Sección 5.2.1) y sus correspondientes validaciones manuales (Véase el Capítulo 8), se creó una nueva versión del léxico, AnCora-Nom-v3, que incluía todas las nominalizaciones deverbales del corpus AnCora-Es, 1.655 lemas, con información sobre la estructura argumental y la denotación (Véase el Capítulo 9). A partir de esta nueva versión del léxico, del corpus AnCora-Es enriquecido con la anotación de las nominalizaciones deverbales (AnCora-Es-v3) y del léxico verbal AnCora-Verb, se desarrolló la versión final del clasificador ADN (ADN-Classifier-v3). A continuación describimos las características técnicas de ADN-Classifier-v3 (Sección 6.1) y después nos centramos en los rasgos utilizados para la clasificación obtenidos de cada uno de los recursos mencionados (Sección 6.2). Cerramos el capítulo con

unas conclusiones (Sección 6.3).

6.1. Clasificador ADN

Definición de la tarea

El objetivo inicial de este clasificador era disponer de una herramienta que nos ayudara a evaluar empíricamente nuestras observaciones sobre la denotación de las nominalizaciones deverbales, tal y como se ha visto en el capítulo anterior. A partir de estos experimentos, nos centramos en un objetivo más ambicioso: proporcionar a la comunidad científica una herramienta de clasificación automática de nominalizaciones deverbales según la denotación capaz de aprovechar la información lingüística disponible en un determinado escenario. Concretamente, el objetivo de ADN es clasificar de forma automática un lema candidato a ser una nominalización deverbal en evento, resultado o subespecificado, así como detectar si dicha nominalización forma parte de una construcción lexicalizada. Como es típico en otras tareas de clasificación léxica, como es el etiquetado morfológico o la desambiguación de sentidos, una palabra tomada como una unidad individual es normalmente ambigua, pero puede ser desambiguada si se tiene en cuenta la información contextual o, al menos, dicha ambigüedad se puede mitigar. Un requisito que deben cumplir las nominalizaciones candidatas es que deben estar etiquetadas morfológicamente como nombres comunes (NC, siguiendo la terminología Parole). De hecho, para llevar a cabo la tarea de clasificación, el clasificador ADN necesita al menos cuatro pre-procesos: 1) tokenización (segmentación del texto a nivel de unidades gráficas o tipográficas); 2) segmentación a nivel oracional; 3) anotación morfosintáctica (PoS); y 4) localización de las nominalizaciones candidatas mediante el uso de una serie de expresiones regulares que buscan sustantivos terminados en 10 sufijos específicos¹.

Pre-procesos necesarios

El clasificador ADN funciona en dos modos: modo aprendizaje, en el que el clasificador aprende el modelo, y modo clasificación, en el que el clasificador aplica el modelo aprendido. La Figura 6.1 esquematiza estos dos modos de funcionamiento. El tipo de aprendizaje es supervisado.

ADN: modo aprendizaje

En modo aprendizaje, el clasificador ADN parte de una muestra de aprendizaje en la que cada ejemplo, que puede ser o un sentido, un lema o una ocurrencia en el corpus AnCora-Es-v3, ha sido manualmente validado. A partir de esta muestra se extrae una matriz de características, que incluye el resultado esperable (la supervisión), que constituye la información de entrada del clasificador ADN, cuyo resultado es la construcción del modelo, un árbol de decisión en nuestro caso. En modo clasificación, se usa el modelo aprendido para clasificar nuevos casos. Previamente, para cada uno de estos nuevos casos se extrae un vector de característi-

ADN: modo clasificación

¹Estos sufijos están detallados en la Sección 3.1.

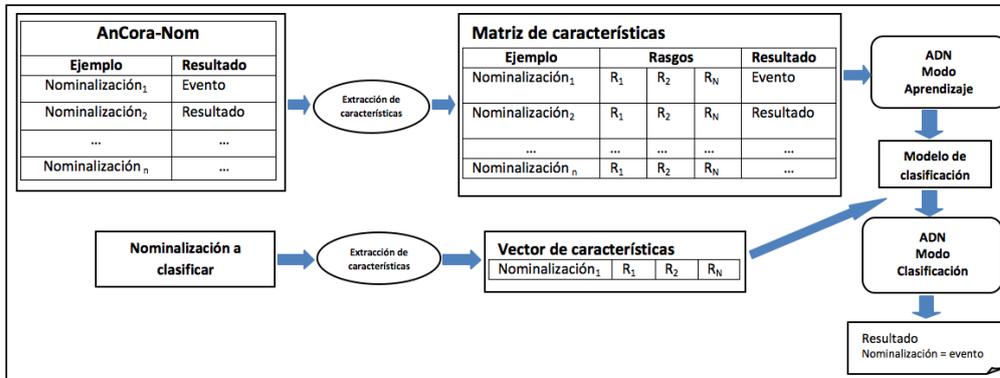


Figura 6.1: Funcionamiento del Clasificador ADN

cas, rasgos que el clasificador ADN usa para clasificar. El vector de características es idéntico al que se da en el modo aprendizaje, no incluyendo, obviamente, el resultado de la clasificación. El resultado del proceso para cada caso es uno de los valores posibles (evento, resultado, subespecificado o construcción lexicalizada) asociado a la nominalización.

De esta manera, un caso para clasificar consiste en un candidato a nominalización y su contexto, esto es, la oración en la que ocurre etiquetada morfológicamente. El candidato y su contexto pueden someterse a procesos complementarios que permiten la extracción de características suplementarias para la clasificación, como son el WSD, el análisis sintáctico, el SRL o relacionar la nominalización candidata con su verbo origen. Cada uno de estos procesos incrementa el número de rasgos para la tarea de clasificación, lo que en principio supondría una mejora en la corrección del clasificador. Sin embargo, todos estos procesos tienen, al mismo tiempo, un coste ya que implican una disminución de la corrección porque el preproceso automático, como es esperable, no es 100 % correcto. Es decir, la actuación de cada módulo incrementa la tasa de error del pre-proceso. Por lo tanto, es necesario un análisis detallado de estos preprocesos, su corrección y su impacto en la corrección de la tarea de clasificación. Por ejemplo, si se usa un sistema de WSD sobre la nominalización candidata, entonces es posible utilizar rasgos del léxico a nivel de sentido en la tarea de clasificación, lo que claramente da lugar a una mejora de la corrección. El inconveniente, sin embargo, es que los resultados actuales en la tarea de WSD no son muy prometedores. En competiciones recientes de SemEval, en la tarea de desambiguación *All-Words* ('todas las palabras') la corrección se sitúa entre el 60 % y el 70 % cuando el caso base, que usa el primer *synset* de WordNet, logra un 51,4 %, mientras que en la tarea de desambiguación *Lexical-Sample* ('muestra léxica') la corrección es de un 89 % siendo el caso base de un 78 % (Pradhan et al., 2007; Chklovski and Mihalcea,

Pre-procesos
complementarios

2002; Decadt et al., 2004). Obviamente, estas cifras dependen del inventario de sentidos (*tagset*) usados para la desambiguación: la tarea *All-Words* usa sentidos más finos (*synsets* de WordNet) mientras que la tarea *Lexical Sample* usa sentidos más groseros (sentidos de Ontonotes).

En la Tabla 6.1 se presentan algunas métricas para describir el corpus AnCora-Es (filas): el grado de polisemia (el número de sentidos por lema), el número de ejemplos, es decir, oraciones, por lema en el corpus, y la longitud media de las oraciones por lema. En las columnas se muestran los valores mínimo y máximo, la media y la desviación estándar para cada una de las métricas.

	Min	Max	Media	Desviación estándar
sentido/lema	1	13	1,86	1,31
ejemplos/lema	1	255	14,15	26,44
longitud oraciones	4	149	39,51	12,08

Tabla 6.1: Contenido descriptivo de AnCora-Es.

Los valores presentados parecen razonables. El único valor anómalo corresponde a la cifra extremadamente alta de desviación estándar en el número de oraciones por lema. Esto se debe al sesgo de la distribución de los valores de esta medida hacia pequeños valores. De hecho, la mayoría de los lemas tienen un ejemplo por lema (415 en total²) y el número de lemas que tienen valores por encima de la media son pocos.

Escenarios

Por lo tanto, nos aproximamos al problema de la clasificación teniendo en cuenta conjuntos diferentes de rasgos provenientes de diferentes recursos lingüísticos (AnCora-Nom, AnCora-Es, AnCora-Verb, etc.) y examinamos y evaluamos la tarea teniendo en cuenta la disponibilidad y conveniencia de las fuentes de conocimiento. Dependiendo de los recursos disponibles y de los procesadores automáticos que se usarían, hemos abordado la tarea de clasificación en diferentes escenarios posibles, que están representados en la Tabla 6.2. Las columnas nos informan de los recursos utilizados en cada escenario (columna 2), de si los rasgos del léxico se extraen a nivel de sentido o a nivel de lema (columna 3) y de los preprocesos automáticos necesarios en cada caso (columna 4).

El escenario 1 en la Tabla 6.2 presenta el caso en el que el léxico nominal (en nuestro caso, AnCora-Nom) estaría disponible y la nominalización candidata es una entrada de dicho léxico. La oración en la que la nominalización se encuentra está etiquetada morfológicamente (PoS) y ningún otro preproceso es aplicado. En este caso, usaríamos ADN con un modelo aprendido solo con rasgos del léxico a

²Adicionalmente, se ha computado el ratio de lemas que contienen un único ejemplo.

Escenario	Recursos	Nivel de Rasgos	Pre-Procesos
1	AnCora-Nom+AnCora-Verb	lema	POS
2	AnCora-Nom+AnCora-Verb	sentido	POS+WSD
3	AnCora-Nom+AnCora-Verb	lema	POS+Parsing
4	AnCora-Nom+AnCora-Verb	sentido	POS+WSD+Parsing
5	AnCora-Nom	lema	POS
6	AnCora-Nom	sentido	POS+WSD
7	AnCora-Nom	lema	POS+Parsing
8	AnCora-Nom	sentido	POS+WSD+Parsing
9	-	lema	POS+Parsing
10	-	lema	POS+Parsing+SRL

Tabla 6.2: Escenarios

nivel de lema con una corrección $\text{Corr}_{\text{lema};\text{lex}}$ (los subíndices indican a qué nivel se obtienen los rasgos -lema o sentido- y de qué recurso provienen). El escenario 2 es el mismo caso que el escenario 1 pero añadiendo un proceso de WSD sobre la nominalización candidata, que obtendría una corrección Corr_{WSD} . En este caso aplicaríamos ADN con un modelo aprendido solo con rasgos del léxico pero a nivel de sentido logrando una corrección $\text{Corr}_{\text{sentido};\text{lex}}$. Obviamente, aplicar este modelo resultaría útil solamente si la $\text{Corr}_{\text{lema};\text{lex}} - \text{Corr}_{\text{sentido};\text{lex}}$ superara el error esperado del WSD ($1 - \text{Corr}_{\text{WSD}}$). El escenario 3 es lo mismo que el escenario 1 pero añadiendo análisis sintáctico de constituyentes (*parsing*) con una corrección $\text{Corr}_{\text{parser}}$. En este caso, aplicaríamos ADN con un modelo aprendido con rasgos del léxico y del corpus a nivel de lema con una corrección $\text{Corr}_{\text{lema};\text{lex}+\text{corpus}}$. Como antes, este modelo resultaría solo útil si la $\text{Corr}_{\text{lema};\text{lex}+\text{corpus}} - \text{Corr}_{\text{lema};\text{lex}}$ mejorara el error esperado del análisis sintáctico ($1 - \text{Corr}_{\text{parser}}$). El escenario 4 es una combinación de los escenarios 2 y 3. Los escenarios 5, 6, 7 y 8 reproducen los escenarios 1, 2, 3 y 4 respectivamente, pero sin usar los rasgos extraídos del léxico verbal AnCora-Verb, por lo que en estos casos no sería necesario establecer la relación de la nominalización candidata con su base verbal. En el escenario 9 el léxico nominal no estaría disponible o la nominalización candidata sería un sustantivo que no está presente en el mencionado lexicón, por lo que solo se usarían los rasgos obtenidos del árbol sintáctico. Finalmente, el escenario 10 sería el mismo que el 9 pero añadiendo un proceso de etiquetado semántico de roles que tendría como objetivo obtener información sobre la estructura argumental paliando, hasta cierto punto, la ausencia de información proveniente del léxico.

Para la construcción de la versión final del clasificador ADN, hemos usado técnicas de ML. En concreto, usamos un clasificador basado en reglas J48.Part –la versión de reglas del clasificador en formato de árbol de decisión C4.5 (Quinlan,

Clasificador: J48.Part

1993)– tal y como está implementado en el programa Weka (Witten and Frank, 2005). Hemos elegido un clasificador en formato de reglas porque nos proporciona una representación natural de las reglas de clasificación, favoreciendo el análisis de cada uno de los modelos sin una disminución en la corrección. Además nos permite realizar un ranking de las diferentes reglas individuales y definir un mecanismo con un umbral con el objetivo de realizar una clasificación orientada a la precisión. Si usáramos clasificadores más complejos basados en reglas como es el Ripper de Cohen (Cohen, 1995), dado que la mayoría de los rasgos son binarios y los discretos tienen un rango pequeño de valores, esto no supondría una mejora respecto a nuestra elección³. A continuación, mostramos un ejemplo simplificado de las reglas generadas por el clasificador J48.Part:

```

...
lex_s.a != argM.adv AND
...
lex_sp = argM.en_loc : argM.durante_tmp : CN:
argM.para_tmp: arg0.contra.agt : argM.por.cau:
argM.en_tmp: argM.con.el.fin.de.fin: argM.a.cau:
arg1.de.pat: argM.de.loc: argM.para.fin
evento (21.0/3.0)

```

Cada regla se compone de diferentes condiciones. Cada condición está formada por el nombre del atributo (*lex_s.a* o *lex_sp*, en el ejemplo anterior), un operador, que puede ser ‘!=’ (no es igual a) y ‘=’ (es igual a), y el valor que se debe o no cumplir (*argM.adv*, *arg1.de.pat*, *argM.para.fin*, etc.). Al final de la regla se proporciona el resultado de clasificación, y entre paréntesis el número de ejemplos positivos y el número de ejemplos negativos del corpus de aprendizaje respecto a esa regla. El ejemplo anterior se interpreta de la siguiente manera: si en el ejemplo a clasificar el atributo del léxico ‘s.a’ no es (‘!=’) *argM.adv* y el atributo del léxico ‘sp’ responde a algunos de los valores indicados (los dos puntos indican disyunción), entonces el resultado de la clasificación es **evento** en 21 casos de 24 (3 ejemplos con estos atributos y valores no responden a esta clasificación).

ADN-Classifier

El clasificador ADN, por lo tanto, consiste en el clasificador J48.Part, tal y como está implementado en el programa Weka, un modelo de clasificación de entre todos los aprendidos (Véase la Sección 7.2 y concretamente la Tabla 7.1) y la lista de rasgos extraídos para la clasificación. De hecho, ADN en modo clasificación

³J48.Part aprende primero el árbol de decisión y después construye las reglas recorriendo las ramas del árbol. Ripper, en cambio, aprende las reglas una por una, lo que hace aumentar considerablemente el coste de aprendizaje. Esto solo podría dar como resultado un conjunto de reglas más pequeño y preciso en el caso de dividir atributos numéricos, que no es nuestro caso.

tiene como información de entrada al sistema una tabla en la que cada fila contiene en la primera columna la nominalización a clasificar y en el resto de columnas los valores para los diferentes rasgos.

6.2. Rasgos utilizados y recursos lingüísticos

En esta sección describimos los rasgos utilizados por el clasificador ADN (en la versión final, -v3) para realizar la tarea de clasificación y los tres recursos lingüísticos (AnCora-Nom, AnCora-Es y AnCora-Verb) de los que provienen dichos rasgos.

6.2.1. Rasgos obtenidos de AnCora-Nom

Los rasgos obtenidos del léxico AnCora-Nom-v3 son los que primero se incorporaron al clasificador como rasgos para el aprendizaje e incluyen los criterios que se revelaron como determinantes en el estudio empírico para establecer la distinción de evento y resultado. A continuación, describimos cada uno de los rasgos que utiliza el clasificador:

a) *Cousin*. Rasgo binario que indica si la nominalización deverbal es o no del tipo *cousin*. Los sustantivos *cousin* son aquellos que no derivan morfológicamente de un verbo, sino que solo mantienen una relación semántica con ellos (‘escarnio-mofarse’) o que dan lugar al verbo (‘revolución-revolucionar’).

b) *Lexicalización*. Este rasgo indica si la nominalización forma parte o no de una construcción lexicalizada (‘golpe de estado’). En caso afirmativo, también se marca el tipo de construcción lexicalizada (nominal, verbal, adjetival, adverbial, preposicional o conjuntiva).

c) *Sentido*. El sentido concreto del lema verbal con el que se relaciona la nominalización también constituye un rasgo para el aprendizaje. Es decir, si una nominalización como ‘distinción’ se deriva del sentido de ‘distinguir’ como sinónimo de ‘diferenciar’ o como sinónimo de ‘destacar’.

d) *Frame*. Cada sentido verbal está asociado a diferentes *frames* o alternancias de diátesis. También se tiene en cuenta el *frame* específico verbal del verbo base de la nominalización, es decir, la estructura sintáctica concreta (transitiva, inacusativa, etc.) de la que deriva la nominalización.

e) *Estructura Argumental*. La estructura argumental de las nominalizaciones es un rasgo complejo que incluye todos los argumentos posibles de una nominalización deverbal. Para cada argumento se detalla también el rol semántico asociado y el constituyente o constituyentes del SN de núcleo deverbal que lo pueden realizar.

f) Complementos no argumentales. También se tienen en cuenta aquellos constituyentes del SN que no son argumentos y que solo modifican al sustantivo.

g) Los tipos de determinantes de la nominalización deverbal también es un rasgo utilizado por ADN.

h) La capacidad de pluralización también es un rasgo de AnCora-Nom para el aprendizaje.

Estos ocho rasgos se obtienen del léxico AnCora-Nom-v3. En el Capítulo 9 se puede ver como están codificados en este léxico con más detalle. A partir de estos diez atributos se han realizado diferentes agrupaciones que afectan especialmente al atributo de la estructura argumental y al de los determinantes. Además de los rasgos agrupados, a partir de los pares atributos-valor del léxico se han generado rasgos binarizados.

6.2.2. Rasgos obtenidos del corpus AnCora-Es

Para la extracción de los rasgos del corpus usamos la herramienta Tgrep2⁴ que nos ha permitido inspeccionar eficientemente los árboles sintácticos en los que está estructurado el corpus AnCora-Es.

Del corpus AnCora-Es-v3 obtenemos básicamente dos tipos de rasgos:

a) Los rasgos contextuales (véase la Tabla 5.6 del capítulo anterior) como son la clase semántica y el tiempo verbal del verbo que domina a la nominalización en la oración, la función sintáctica del SN cuyo núcleo es la nominalización o si la nominalización aparece en un SN que sea o no una entidad con nombre. Recuérdese que necesitábamos de 108 reglas tgrep para obtener estos rasgos.

b) Además, del corpus también se obtiene la versión del corpus de algunos de los rasgos del léxico, es decir, rasgos que se pueden obtener del corpus si no se dispone del léxico, como el tipo de especificador, si el sustantivo puede aparecer en plural o no y los constituyentes que son complementos de las nominalizaciones y el tipo de preposición en el caso de los SPs. Estos rasgos son especialmente útiles en los modelos en los que se prescinde de los rasgos del léxico. En total, se han construido 72 reglas tgrep adicionales para extraer estos rasgos⁵.

6.2.3. Rasgos obtenidos del léxico AnCora-Verb

Otros de los recursos que utiliza el clasificador ADN es el léxico verbal AnCora-Verb-Es, que contiene 2.830 verbos del español asociados a distintas clases semánticas. La clase semántica del verbo base de la nominalización se usa como rasgo

⁴<http://tedlab.mit.edu/~dr/TGrep2/>. Tgrep2 es una mejora de la herramienta Tgrep. Ambas son herramientas aplicadas a árboles sintácticos basadas en la ampliamente utilizada herramienta Unix Grep sobre cadenas de caracteres.

⁵Estas reglas también se pueden consultar en el link: http://clic.ub.edu/corpus/webfm_send/61



Figura 6.2: Árbol sintáctico parcial que contiene la nominalización ‘aumento’

para el aprendizaje en el clasificador ADN, es decir, se tiene en cuenta si la nominalización deriva de un verbo de la clase de las realizaciones, los logros, los estados o las actividades, ya que puede influir en el tipo denotativo de la nominalización.

6.3. Conclusiones

En este capítulo se ha presentado el clasificador ADN (-v3), sus características técnicas y los recursos de los que obtiene los rasgos para el aprendizaje. Aunque ADN nació a partir de los experimentos computacionales realizados para eva-

luar la información manualmente anotada en AnCora-Nom-v1 (Véase el capítulo anterior), ha llegado a convertirse en una herramienta de clasificación de nominalizaciones deverbales del español capaz de actuar en diferentes escenarios. En el siguiente capítulo se detallan los experimentos realizados con la versión final de este clasificador, su evaluación, así como la evaluación de la eficiencia de ADN en los diferentes escenarios.

CAPÍTULO 7

CLASIFICADOR ADN: EXPERIMENTOS

En este capítulo se presentan los experimentos realizados para construir nuevos modelos de clasificación (a nivel de sentido y a nivel de lema) a partir de los recursos creados, es decir, AnCora-Nom-v3 y AnCora-Es-v3, unos modelos que aprenden con un mayor número de instancias y con recursos totalmente validados. Además también se han replicado los experimentos a nivel de sentido y lema realizados en los primeros experimentos con ADN (Sección 5.2.1) con el subconjunto de 100.000 palabras de AnCora-Es-v3 y el subconjunto de 817 entradas léxicas de AnCora-Nom-v3. Para la evaluación de todos estos nuevos modelos desarrollados, basados en sentidos y basados en lema, se ha utilizado la validación cruzada con 10 particiones a partir de AnCora-Nom-v3 y AnCora-Es-v3. Estos modelos dan lugar a la versión final del clasificador ADN (-v3). En primer lugar, presentamos el marco de desarrollo de estos experimentos (Sección 7.1), en segundo lugar, nos centramos en la descripción de los experimentos (Sección 7.2) para luego evaluar los resultados (Sección 7.3) y comparar nuestro trabajo con otros trabajos similares (Sección 7.4). Finalmente terminamos con unas conclusiones (Sección 7.5).

7.1. Marco de desarrollo

Con el objetivo de evaluar la eficiencia de ADN, se diseñaron dos series de experimentos. Por una parte, se experimentó con diferentes modelos del clasificador que se parametrizan en cinco dimensiones (Veáse la Sección 7.2) y por otra, de los diferentes modelos creados, aplicamos los más adecuados a los escenarios resumidos en la Tabla 6.2 de la Sección 6.1. Para la evaluación de estas dos series de experimentos se utilizó la validación cruzada con 10 particiones a partir de AnCora-Nom-v3 y AnCora-Es-v3.

Para la evaluación de los rasgos seleccionados y para llevar a cabo la tarea de clasificación en cada escenario, hemos usado los modelos descritos en la Sección 7.2. Como ya se ha dicho, usar ADN para la clasificación en una tarea completa implica usar el clasificador basado en reglas J48.Part, tal y como está implementado en el programa Weka, y el modelo de clasificación adecuado para cada caso de entre todos los aprendidos.

7.2. Experimentos

En esta sección se describen los experimentos realizados con el clasificador ADN. Primero, nos centramos en aquellos experimentos relacionados con los diferentes modelos generados y a continuación discutimos cómo algunos de esos modelos se aplican en los diferentes escenarios (Peris et al., 2012).

Los diferentes modelos de ADN aplicados se estructuran en las siguientes cinco dimensiones:

- **Nivel de aplicación.** Distinguimos entre modelos a nivel de sentido y modelos a nivel de lema. Los primeros usan la información del léxico AnCora-Nom a nivel de sentido, esto es, asociando los rasgos para el aprendizaje y clasificación a los diferentes sentidos de la nominalización. En cambio, en los modelos a nivel de lema, a la hora de extraer los rasgos del léxico, se usan como rasgos para el aprendizaje y la clasificación aquellos atributos cuyos valores son compartidos por todos los sentidos de un mismo lema. Por lo tanto, en este segundo nivel de aplicación los rasgos obtenidos del léxico no son tan informativos pero al mismo tiempo se reduce la dependencia del léxico AnCora-Nom, lo que constituye un paso necesario para convertir la tarea de clasificación en más realista. Los modelos a nivel de sentido tienen más interés para el estudio teórico del fenómeno de la denotación en las nominalizaciones deverbales, mientras que los modelos a nivel de lema tienen más interés para la clasificación automática en un escenario real en el que, posiblemente, la selección del sentido correcto no es factible.
- **Unidad de aprendizaje y clasificación** (la instancia a ser clasificada). Tanto los modelos a nivel de sentido como los modelos a nivel de lema son a su vez distinguidos según si la unidad de clasificación procede del léxico –sentido o lema– o del corpus AnCora-Es –ejemplos. En el primer caso, la unidad de clasificación agrupa a todas las oraciones del corpus que correspondan al lema o sentido considerado; en el segundo caso, se considera a cada una de las oraciones por separado. En consecuencia en los modelos a nivel de sentido las unidades utilizadas son sentidos o ejemplos del corpus, mientras que en los modelos a nivel de lema las unidades son los lemas o los

ejemplos del corpus. Cuando las unidades son sentidos o lemas los rasgos usados en los correspondientes modelos se obtienen solo de AnCora-Nom, mientras que si la unidad son los ejemplos también podemos usar los rasgos contextuales obtenidos del corpus. Se tiene que tener en cuenta que según esta dimensión el número de instancias para el aprendizaje varía ya que hay más sentidos que lemas en el léxico y porque las ocurrencias de nominalizaciones en el corpus (ejemplos) son más que los lemas o los sentidos del léxico.

- **Rasgos implicados.** Los rasgos utilizados para el aprendizaje y la clasificación se obtienen del léxico (rasgos léxicos) o del corpus (rasgos contextuales). Los diferentes modelos se pueden distinguir también por utilizar exclusivamente rasgos léxicos, rasgos contextuales o por una combinación de ambos tipos de rasgos.
- **Tamaño del léxico.** Los conjuntos de datos tenidos en cuenta se corresponden con el conjunto reducido de 817 lemas obtenidos del subconjunto de 100.000 palabras del corpus AnCora-Es usados en los primeros experimentos del ADN, o el conjunto completo de 1.655 lemas deverbales de AnCora-Nom obtenidos del corpus AnCora-Es en su totalidad (500.000 palabras). Dependiendo de esta dimensión el número de instancias para el aprendizaje también varía.
- **Tamaño del corpus.** Dos conjuntos del corpus se usan en los diferentes modelos: el subconjunto de 100.000 palabras usado también en los primeros experimentos y el AnCora-Es en su totalidad (500.000 palabras)¹. Dependiendo de esta dimensión el número de instancias para el aprendizaje también varía.

Para identificar los modelos que se presentan en la Tabla 7.1, usamos un código de cinco letras como notación y cada una de ellas identifica una de las 5 dimensiones mencionadas. La primera letra corresponde al nivel de aplicación: si el modelo es a nivel de sentido usaremos una S para su identificación y una L en el caso de los modelos a nivel de lema. La segunda letra se refiere a la unidad de clasificación y es L para los lemas, S para los sentidos y E para los ejemplos. En la tercera posición encontramos la referencia a la fuente a partir de la cual se obtienen los rasgos: L para el léxico, C para el corpus y A para la combinación de ambos tipos de rasgos. En cuarto lugar nos referimos al tamaño del léxico y la letra R representa al conjunto reducido de 817 lemas y la F corresponde al conjunto

Modelos

¹Para obtener la curva de aprendizaje de alguno de nuestros modelos se han utilizado tamaños intermedios del corpus.

completo de 1.655 lemas. En último lugar, también designamos el tamaño del corpus con R en caso del subconjunto de 100.000 palabras y con un F para el corpus completo. Por ejemplo, un modelo a nivel de sentido que usa ejemplos como unidad de clasificación, utiliza rasgos del léxico y del corpus y usa el léxico y corpus al completo se identifica como el modelo LEAFF. En total hemos experimentado con 32 modelos diferentes².

Para los diferentes escenarios presentados en la Sección 6.1 del Capítulo 6, aplicamos los modelos más apropiados de acuerdo a la información que posee cada escenario de los modelos creados, que pueden utilizar todos los rasgos del modelo o bien podemos excluir algunos rasgos en alguno de los modelos para que se ajuste mejor a cada escenario.

7.3. Evaluación

Para evaluar la eficacia de los diferentes modelos, evaluamos cada uno de ellos mediante el método de validación cruzada con 10 particiones. A continuación nos centramos en los resultados de los 32 modelos resultantes de las cinco dimensiones descritas en la Sección 7.2. Luego presentamos los resultados de uno de los modelos orientado a la precisión (Subsección 7.3.1), la evaluación de los escenarios descritos en la Tabla 6.2 (Subsección 7.3.2) y un análisis de errores (Subsección 7.3.3).

Resultados

La Tabla 7.1 presenta los resultados obtenidos por los 32 modelos: en la columna 1 se presentan mediante la anotación antes descrita los 32 modelos, el número de ejemplares usados para el aprendizaje se detalla en la columna 2, el número de rasgos utilizados y las reglas generadas por el clasificador se muestran en las columnas 3 y 4, y finalmente encontramos el caso base, la corrección, el decrecimiento del error sobre el caso base (Δ -error) y el ratio relativo de reducción de error (Red- Δ -error) obtenidos en cada modelo en las columnas 5, 6, 7 y 8, respectivamente. Las filas se corresponden con los diferentes modelos presentados. Cabe notar que en la columna 2, el número de ejemplares para el aprendizaje depende del tipo de unidad usada para el aprendizaje y la clasificación (sentidos, lemas y ejemplos) y del tamaño del léxico y del corpus. La interacción de estas tres dimensiones también explica por qué las cifras del caso base cambian en cada modelo. El caso base asigna a todas las instancias la clase más frecuente, es decir, la clase resultado. En general, cuando la unidad utilizada se obtiene del léxico el caso base del lema supera al del sentido. Esto se explica porque en los modelos a nivel de lema agrupamos bajo un lema los sentidos que comparten todos los

²No todas las combinaciones entre las cinco dimensiones están permitidas, por ejemplo, los modelos a nivel de sentido excluyen los lemas como unidad de aprendizaje y a la inversa. Es por esta razón que los modelos son 32 y no 72 ($2 \times 3 \times 3 \times 2 \times 2$).

rasgos; dado que es infrecuente que diferentes sentidos compartan toda la información, a efectos prácticos, solo los lemas monosémicos son tenidos en cuenta. Este hecho, por tanto, muestra que hay más lemas monosémicos de tipo resultativo que de tipo eventivo y subespecificado. Además, es importante remarcar que cuando la unidad de aprendizaje y clasificación utilizada son los ejemplos del corpus y no los sentidos del léxico, el caso base también aumenta. De hecho, las nominalizaciones resultativas están más ampliamente representadas en el corpus que las eventivas y las subespecificadas. En lo que respecta al número de rasgos usados en el aprendizaje, las dimensiones relevantes son el tipo de rasgo y el tamaño del léxico (cuando se usan rasgos de este recurso). Finalmente, cabe decir que la corrección y las otras dos medidas relacionadas se obtienen tras la evaluación de la eficiencia del clasificador mediante el método de la validación cruzada con 10 particiones a partir de AnCora-Nom-v3 y AnCora-Es-v3.

Como se observa en la Tabla 7.1, los modelos a nivel de sentido (las 16 primeras filas) funcionan mejor que los correspondientes modelos a nivel de lema (las 16 últimas filas). Esto se explica por el hecho de que los rasgos del léxico a nivel de sentido no pueden ser recuperados a nivel de lema ya que en los modelos de lema solo se usan los rasgos que son compartidos por todos los sentidos del mismo lema y esto no es muy frecuente, es decir, buena parte de los rasgos no están informados y ello produce una caída en la corrección del clasificador aprendido. A nivel de sentido los mejores resultados se obtienen si los rasgos utilizados en la clasificación proceden exclusivamente de AnCora-Nom, siendo la unidad de clasificación el sentido del léxico (el primer bloque de cuatro filas) o los ejemplos del corpus (el segundo bloque de cuatro filas). Los rasgos contextuales (aquellos que se extraen del corpus) perjudican la corrección: cuando se usan solos (el tercer bloque de cuatro filas), los valores de corrección están por debajo del caso base y cuando se usan en combinación con los rasgos obtenidos del léxico, la corrección disminuye si la comparamos con los modelos que solo usan el léxico como fuente de obtención de rasgos. Esto demuestra que el léxico contiene información crucial que no es posible recuperar del corpus. Además, se debe mencionar que de forma general en los modelos de nivel de sentido existe una mejora generalizada que se explica por el tamaño del léxico y especialmente del corpus: cuánto más grande es el vocabulario y el corpus usado, mejor es el resultado. Este hecho también está presente en los modelos a nivel de lema.

Los modelos a nivel de sentido representan la cota superior de nuestra tarea. Sin embargo, en un marco más realista y teniendo en cuenta el estado de la cuestión en WSD, no tendríamos acceso a las etiquetas de sentido por lo que estamos mucho más interesados en la eficiencia de ADN en los modelos a nivel de lema. Los mejores resultados se logran cuando se combinan rasgos del léxico y del corpus (último bloque de cuatro filas), lo que muestra que la suma de ambos tipos de rasgos da lugar a resultados positivos, lo que no consiguen ni los rasgos del léxico

Resultados: modelos a nivel de sentido

Resultados: modelos a nivel de lema

	Modelo	Inst.	Atr.	Reglas	Caso Base	Corr.	Δ error	Red Δ error	
Modelos a nivel de sentido	Sentidos	SSLRR	609	937	51	71,75 %	76,84 %	5,09 %	18,02 %
		SSLRF	964	937	78	60,68 %	70,02 %	9,33 %	23,74 %
		SSLFR	1.428	1.671	84	70,86 %	81,72 %	10,85 %	37,25 %
		SSLFF	3.094	1.671	224	60,95 %	74,36 %	13,41 %	34,35 %
	Ejemplos	SELRR	1.840	937	42	85,32 %	93,80 %	8,47 %	57,77 %
		SELRF	9.278	937	137	87,03 %	97,82 %	10,78 %	83,20 %
		SELFR	3.994	1.671	117	83,92 %	93,69 %	9,76 %	60,74 %
		SELFF	23.431	1.671	366	85,45 %	96,65 %	11,19 %	76,99 %
		SECRR	1.840	197	35	85,32 %	83,96 %	-1,35 %	-9,25 %
		SECRF	9.278	197	116	87,03 %	86,34 %	-0,68 %	-5,31 %
		SECFR	3.994	197	81	83,92 %	82,72 %	-1,20 %	-7,47 %
		SECFE	23.431	197	211	85,45 %	84,93 %	-0,52 %	-3,60 %
	Lemas	LEARR	1.840	1.133	76	85,32 %	91,57 %	6,25 %	42,59 %
		LEARF	9.278	1.133	196	87,03 %	96,38 %	9,35 %	72,15 %
		LEAFR	3.994	1.867	146	83,92 %	91,72 %	7,80 %	48,52 %
		LEAFF	23.431	1.867	498	85,45 %	95,46 %	10,01 %	68,83 %
Modelos a nivel de lema	Lemas	LLLRR	242	852	6	90,90 %	88,84 %	-2,06 %	-22,72 %
		LLLRF	242	852	6	90,90 %	88,84 %	-2,06 %	-22,72 %
		LLLFR	532	1.559	14	89,84 %	89,66 %	-0,18 %	-1,85 %
		LLLFF	972	1.559	26	87,55 %	89,09 %	1,54 %	12,39 %
	Ejemplos	LELRR	1.840	852	50	85,32 %	83,96 %	-1,35 %	-9,25 %
		LELRF	9.278	852	76	87,03 %	86,88 %	-0,15 %	-1,16 %
		LELFR	3.994	1.559	162	83,92 %	83,50 %	-0,42 %	-2,64 %
		LELFF	23.431	1.559	322	85,45 %	85,62 %	0,16 %	1,14 %
		LEARR	1.840	197	35	85,32 %	84,02 %	-1,30 %	-8,88 %
		LEARF	9.278	197	116	87,03 %	86,35 %	-0,67 %	-5,23 %
		LECFR	3.994	197	81	83,92 %	82,57 %	-1,35 %	-8,41 %
		LECFE	23.431	197	211	85,45 %	84,86 %	-0,58 %	-4,04 %
	Ejemplos	LEARR	1.840	1.048	109	85,32 %	85,05 %	-0,27 %	-1,85 %
		LEARF	9.278	1.048	355	87,03 %	87,64 %	0,61 %	4,7 %
		LEAFR	3.994	1.755	236	83,92 %	85,27 %	1,35 %	8,41 %
		LEAFF	23.431	1.755	981	85,45 %	87,20 %	1,74 %	12,00 %

Tabla 7.1: Experimentos y Evaluación de los modelos. Leyenda de los nombres de los modelos: 1ª letra (S= nivel de sentido, L= nivel de lema); 2ª letra (S= sentido, L= lema, E= ejemplo del corpus); 3ª letra (L= rasgos del léxico, C= rasgos del corpus, A= ambos tipos de rasgos); 4ª letra (R= vocabulario reducido, F= vocabulario entero) y 5ª letra (R= corpus reducido, F= corpus entero)

ni los rasgos del corpus por separado. Cuando los rasgos usados en la clasificación se obtienen exclusivamente del léxico, siendo la unidad de clasificación los lemas del léxico (el quinto bloque de cuatro filas) o los ejemplos del corpus (el sexto bloque de cuatro filas), los resultados son siempre negativos (por debajo del caso

base) excepto cuando el tamaño del léxico y corpus son ambos completos (una mejora del 1,54 % y 0,16 %, respectivamente). En estos casos la información del léxico no es tan específica como en los modelos a nivel de sentido. Los rasgos del corpus por sí solos tampoco logran resultados positivos ni siquiera cuando el tamaño del léxico y del corpus son completos. Por lo tanto, la combinación de rasgos del léxico y del corpus es necesario en un marco más realista para lograr un buen resultado del clasificador. En los casos combinados, solo cuando se usa el tamaño del léxico y del corpus reducidos los resultados son ligeramente negativos. A partir de ahora centraremos nuestra atención en el último modelo (LEAFF) porque aunque obtenga una corrección más baja que el correspondiente modelo a nivel de sentido, se espera que este modelo muestre un comportamiento más robusto a la hora de enfrentarse a datos no conocidos.

Un aspecto importante para que el clasificador aprenda un modelo es si la muestra de aprendizaje es o no suficientemente amplia para un aprendizaje correcto. Hemos llevado a cabo un análisis de la curva de aprendizaje del modelo LEAFF teniendo en cuenta diferentes tamaños, de porciones del corpus de 1.000 ejemplos a los ejemplos correspondientes a todo el corpus, 23.431. Los resultados obtenidos se muestran en la Figura 7.1. También se muestran los intervalos de confianza al 95 %. Los resultados dejan entrever que para muestras más grandes de 5.000 ejemplos la corrección tiende a estabilizarse, por lo que estamos ampliamente seguros de nuestros resultados. Como era previsible, los intervalos de confianza se hacen más estrechos al aumentar el tamaño del corpus de aprendizaje aunque no dejan de tener una amplitud notable que no disminuye mucho con el tamaño del corpus. Ello puede interpretarse en términos de lo que se puede aprender del corpus ya está aprendido a partir de unos 5.000 ejemplos y que para posibles mejoras haría falta incorporar otros rasgos.

LEAFF, curva de aprendizaje

7.3.1. Clasificador orientado a la precisión

Todos los experimentos presentados hasta ahora se basan en una cobertura completa, es decir, del 100 % y, por lo tanto, la corrección y la precisión tienen el mismo valor. Adicionalmente, hemos llevado a cabo experimentos con el clasificador orientado a la precisión, con el objetivo de lograr una precisión alta a expensas de una caída en la cobertura, es decir, admitiendo que algunos casos queden sin clasificar. Para realizar este objetivo, se ha asignado un valor (la precisión individual de la regla proporcionada por Weka) a cada una de las reglas del modelo LEAFF (981 reglas) sin tener en cuenta el orden en que estas reglas se aplican. Hemos ordenado dichas reglas por sus valores individuales decrecientes y hemos construido un clasificador basado en un mecanismo de umbral: solo las reglas con precisión por encima del umbral se han aplicado. Esto ha dado lugar a una precisión más alta pero ha tenido un coste en la caída de la cobertura. Los

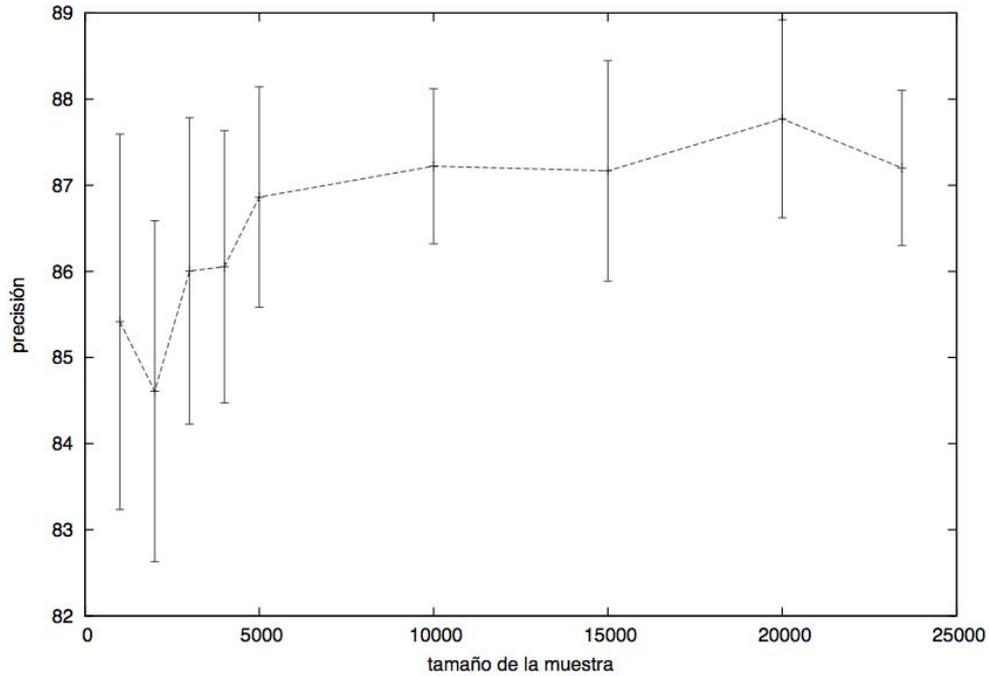


Figura 7.1: Curva de aprendizaje para el modelo LEAFF

resultados se muestran en la Figura 7.2. Como se puede ver, hasta un 68 % de las reglas con un valor bajo se pueden eliminar sin una caída apreciable en la precisión. Además, se logra una precisión del 90 % con solo una caída del 4 % en la cobertura o una precisión del 80 % con solo una caída del 2 % en la cobertura.

7.3.2. Evaluación de los escenarios

Los resultados de los experimentos realizados en los escenarios descritos en la Sección 6.1 (Véase la Tabla 6.2) se presentan en la Tabla 7.2. Esta tabla nos muestra los resultados de los diez escenarios en las diferentes filas de la tabla. En las columnas encontramos la identificación del escenario (columna 1); el modelo aplicado, que sigue la notación descrita en la Sección 7.2 (columna 2); el número de rasgos en el modelo original (columna 3); el número de rasgos del modelo aplicado tras eliminar los rasgos no informados (columna 4); y los valores de corrección del modelo original y del modelo final (columna 5 y 6, respectivamente).

Estos resultados muestran que la diferencia entre los modelos a nivel de sentido y los modelos a nivel de lema mostrada en la Tabla 7.1 también está presente

7. CLASIFICADOR ADN: EXPERIMENTOS

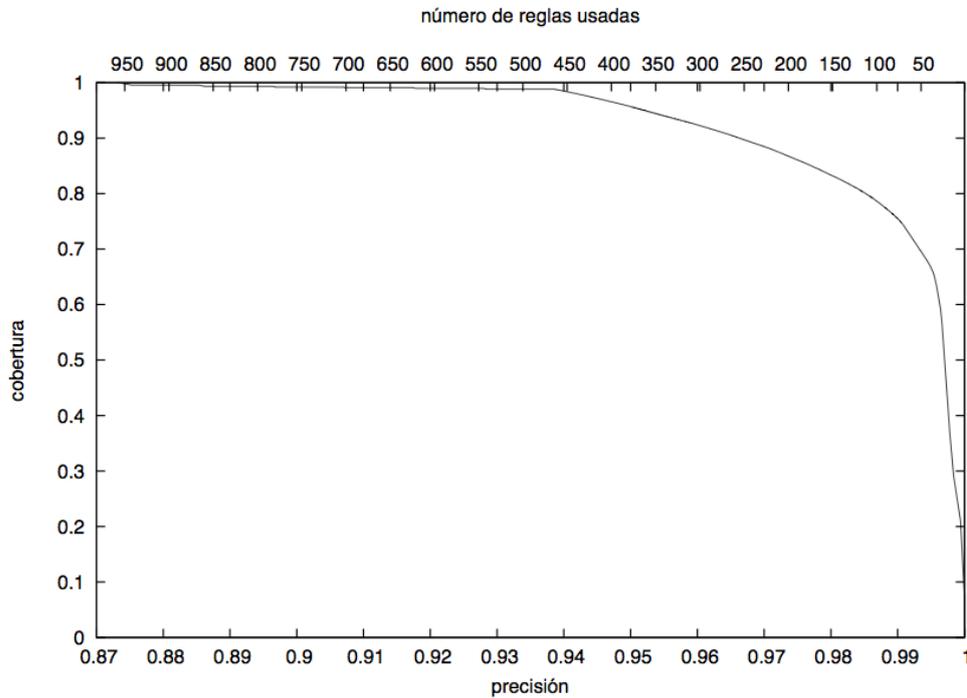


Figura 7.2: Cobertura y precisión para el modelo LEAFF.

Escenario	Modelo	Rasgos. Ini.	Rasgos. Fin.	Corr. Ini.	Corr. Final
1	LELFF	1.559	1.559	85,62 %	85,62 %
2	SELFF	1.671	1.671	96,65 %	96,65 %
3	LEAFF	1.755	1.755	87,20 %	87,20 %
4	SEAFF	1.867	1.867	95,46 %	95,46 %
5	LELFF	1.559	1.416	85,62 %	85,56 %
6	SELFF	1.671	1.613	96,65 %	96,17 %
7	LEAFF	1.755	1.611	87,20 %	87,12 %
8	SEAFF	1.867	1.808	95,46 %	95,41 %
9	LECFE	197	197	84,86 %	84,86 %
10	LEAFF	1.755	1.556	87,20 %	87,08 %

Tabla 7.2: Experimentos y evaluación de los escenarios

en el caso de los escenarios. En todos los escenarios en los que algunos rasgos son eliminados hay una disminución de la corrección, si bien esta no es estadísti-

camente significativa. Esto se debe a que el número de rasgos para aprender es elevado y que el clasificador usa rasgos alternativos cuando se eliminan otros.

7.3.3. Análisis de errores

El análisis de errores se centra en el modelo a nivel de lema que usa tanto rasgos del léxico como del corpus y que aprende con el léxico y el corpus completo (LEAFF). La Tabla 7.3 presenta la matriz de confusión de este modelo. Las filas se corresponden con la validación manual y en las columnas se encuentran las predicciones del clasificador. Las predicciones correctas se encuentran en la diagonal en negrita. Los errores están marcados en cursiva.

Matriz de confusión

Clasificación manual etiqueta correcta ⇒	Clasificación de ADN ↓				Total valores correctos
	R	E	SE	L	
R	18.997	<i>575</i>	<i>397</i>	<i>54</i>	20.023
E	<i>676</i>	933	<i>242</i>	<i>2</i>	1.853
SE	<i>643</i>	<i>309</i>	453	<i>7</i>	1.412
L	<i>90</i>	<i>2</i>	<i>2</i>	49	143
Total ADN	20.406	1.819	1.094	112	23.431

Tabla 7.3: Matriz de confusión del modelo LEAFF. Leyenda: R= resultado; E= evento; SE= subespecificado y L= lexía.

El porcentaje de error se reparte de manera casi igual entre las tres clases principales de denotaciones: las nominalizaciones eventivas incorrectamente clasificadas representan un 31 % de los errores, las resultativas mal clasificadas son el 34 % y las subespecificadas el 32 %. Las nominalizaciones lexicalizadas, sin embargo, solo suponen el 3 % del error³.

Nominalizaciones eventivas

Entre las nominalizaciones eventivas incorrectamente clasificadas por ADN, el 73 % (676 instancias) fueron clasificadas como resultativas, el 26 % (242 instancias) como subespecificadas y un marginal 3 % (2 instancias) como lexías no nominales. Estos errores se explican por cuatro razones principales.

En primer lugar, un 27 % de los errores son, de hecho, errores en la validación manual⁴ (1), lo que significa que ADN las clasificó correctamente. En el ejemplo (1), los anotadores habían etiquetado ‘formación’ como evento, sin embargo, el

³En este caso, nos referimos a nominalizaciones lexicalizadas a las que no se asigna ningún tipo de denotación, es decir, a las lexías no nominales.

⁴En el proceso de comparación de la anotación automática con la validación manual, alguna

adjetivo relacional ‘abertzale’ da la pista que el sustantivo es un resultado, tal y como ADN proponía. Además, la prueba de la paráfrasis, ‘los abertzales forman una formación’, refuerza la lectura resultativa.

En segundo lugar, otro 51 % se explica porque la guía de anotación referente a la denotación contiene ciertos criterios a los que ADN no tiene acceso y los anotadores en la validación manual sí: los llamados selectores y el criterio del agente y de la paráfrasis (Véase la Sección 5.1.2), es decir, se trata de casos que ADN nunca podría clasificar correctamente con el conocimiento que posee. En el ejemplo (2), ADN había clasificado el sustantivo como subespecificado, sin embargo, la preposición ‘durante’ funciona como selector e indica la lectura eventiva.

En tercer lugar, un error del 21 % en la clasificación de eventos se explica porque hay un número de criterios, implementados como rasgos en ADN, que sufren de escasez de datos y, por lo tanto, ADN no los puede aprender como lo relevante que son. Por ejemplo, un dato determinante para detectar nominalizaciones eventivas consiste en un determinante posesivo interpretado como arg1-pat (paciente), pero al no ser muy frecuente ADN no lo aprende como lo relevante que es. Este es el caso del ejemplo (3) que ADN había clasificado como resultado y en realidad es un evento.

Finalmente, los casos en los que ADN clasificó nominalizaciones eventivas como construcciones lexicalizadas se explican porque el clasificador confundió estas ocurrencias con construcciones lexicalizadas que compartían el mismo lema (el porcentaje de error es del 1 %). El ejemplo (4), ADN lo había anotado como lexía porque lo ha confundido con un lexicalización que tiene el mismo lema ‘centro de acogida’.

- (1) Garzón ha admitido el recurso que ha presentado [la **formación**<resultado> [abertzale]_{SA-arg1-tem}]_{SN}.
- (2) Admitieron hoy durante [[su]_{Poss-arg0-agt} **declaración**<evento> [el traspaso del jugador]_{SP-arg1-pat}]_{SN}.
- (3) Dejamos un país tan claro para [[su]_{Poss-arg1-pat} **gobierno**<evento>]_{SN} como el cielo después de la tormenta.
- (4) Se preguntan por [la **acogida**<evento> [de la medida]_{SP-arg1-pat} [por los pensionistas]_{SP-arg0-agt}]_{SN}.

En las nominalizaciones resultativas incorrectamente etiquetadas por el clasificador ADN, un porcentaje del 56 % (575 instancias) se clasificaron como even-

Nominalizaciones
resultativas

de las correcciones manuales se consideraron dudosas. Estos casos se discutieron entre todos los anotadores y se decidió qué anotación (automática o manual) era la correcta. Por lo tanto, nos referimos a aquellos casos finalmente considerados como incorrectamente clasificados en el proceso de validación manual como errores manuales.

tivas; un 39 % (397 instancias) como subespecificadas y un 5 % (54 instancias) como nominalizaciones lexicalizadas. Estos errores se explican por las mismas cuatro razones citadas arriba. El porcentaje del error manual es en este caso del 51 %, lo que significa que existen nominalizaciones eventivas y subespecificadas que estaban incorrectamente validadas. El ejemplo (5), los anotadores habían asignado la etiqueta resultado a ‘privatización’ pero el sustantivo proceso funciona como selector y muestra que es una nominalización eventiva como proponía ADN.

El porcentaje de error explicado por lo selectores es solo del 10 % porque tenemos más selectores para identificar nominalizaciones eventivas que para identificar nominalizaciones resultativas. El ejemplo (6) muestra un caso que ADN había clasificado como evento, pero que en la validación manual se ha anotado como resultado puesto que no hay criterios que indiquen eventividad y porque la combinación con un predicado atributivo induce a la lectura resultativa.

Además, un 37 % del porcentaje de error se explica por aquellos criterios que aunque implementados como rasgos en ADN, sufren de escasez de datos y, por lo tanto, a pesar de su relevancia, ADN no los puede aprender. En el caso de las nominalizaciones resultativas existen más criterios de este tipo: por ejemplo, las nominalizaciones que derivan de verbos de actividades o estativos, el hecho de tener adjetivos relacionales como argumentos, argumentos temporales realizados mediante un SP introducido por la preposición ‘de’. En el ejemplo (7) observamos que ‘naufragio’, que había sido clasificado como subespecificado por ADN es en realidad un resultado, como indica el hecho de que el verbo base sea de la clase D (verbo de actividad) y tenga como argumento temporal un SP introducido por la preposición ‘de’.

Finalmente, los casos en los que ADN clasificó nominalizaciones resultativas como construcciones lexicalizadas se explican porque el clasificador confundió estas ocurrencias con construcciones lexicalizadas que compartían el mismo lema (el porcentaje de error es del 2 %). En el ejemplo (8) la confusión se produce con la lexía ‘estado de excepción’.

- (5) Se comprometió a revisar el proceso de [**privatización**<evento> [de su antecesor]SP-arg0-agt]SN.
- (6) [La **conexión**<resultado> [que tenemos con el Magreb]S-RefMod]SN es insuficiente.
- (7) Quiere arreglar [[su]Poss-arg0-agt **naufragio**<resultado> [del jueves]SP-argM-tmp]SN.
- (8) Todos los modelos están más baratos con [la **excepción**<resultado> [del modelo Peugeot]SP-arg1-pat]SN.

En las nominalizaciones subespecificadas que ADN clasifica incorrectamente, un 32 % (309 instancias) fueron clasificadas como eventivas, un 67 % (643 instancias) como resultativas y un 1 % marginal como construcciones lexicalizadas. Es interesante resaltar que es el único tipo de nominalización en el que el clasificador falla más veces que acierta. De hecho, ADN clasifica como tales a 1.094 casos frente a los 1.412 casos manuales. La dificultad que presentan las nominalizaciones subespecificadas es esperable puesto que se trata de casos en los que no se tiene ningún rasgo contextual que permita la desambiguación o casos que pese a tener contexto son realmente ambiguos. En este caso, el porcentaje de error en la validación manual es del 45 %. En el ejemplo (9), los anotadores habían asignado la etiqueta subespecificado pero han pasado por alto que el verbo del que deriva la nominalización ‘diferencia’ es estativo y, por lo tanto, la nominalización es resultativa, como proponía ADN.

Nominalizaciones
subespecificadas

Aunque no existen selectores que identifiquen a las nominalizaciones subespecificadas (debido a su naturaleza), en algunos casos si el SN presenta criterios contradictorios los anotadores manuales recibieron la instrucción de anotar la nominalización como subespecificada. Por ejemplo, un artículo indefinido indica que la nominalización es resultativa y el selector ‘durante’ selecciona típicamente una nominalización eventiva, si una nominalización presenta estos dos criterios contradictorios entre sí, se anota como subespecificada. Sin embargo, como el clasificador ADN no tiene acceso a los selectores, la mayoría de casos han sido anotados como resultativos. Esto es lo que ocurría en el ejemplo (10), pero los anotadores, ante la interacción entre el selector ‘durante’ y el artículo indefinido, le han asignado la etiqueta subespecificada. Este tipo concreto de error representa el 19 % de las nominalizaciones subespecificadas incorrectamente clasificadas.

- (9) [La gran **diferencia**<resultado> [de mi actitud]_{SP-arg1-tem}]_{SN} es que yo no hablo demasiado.
- (10) Le hicieron las heridas durante [un **interrogatorio**<subespecificado>]_{SN}.
- (11) Es el máximo órgano competente en [la **interpretación**<subespecificado> [de tratados internacionales]_{SP-arg1-pat}]]_{SN}.
- (12) Es lo más suave que se puede decir d[el **desenlace**<subespecificado> [del debate sobre las pensiones]_{SP-arg1-tem}]_{SN}.
- (13) Permanecerá como gerente del grupo hasta [[su]_{Poss-arg0-agt} **jubilación**<subespecificado>]_{SN}.

El criterio del agente explica un 20 % del error global en las nominalizaciones subespecificadas. Si tanto el SP introducido por la preposición ‘por’ como el introducido por la preposición ‘de’ son validos para la nominalización que los anotadores están validando, ellos anotaban dicha nominalización como subespecificada. Por

lo tanto, los anotadores disponen una vez más de un criterio del que ADN no puede hacer uso. En el ejemplo (11), el SN que contiene la nominalización tanto podría admitir un agente introducido por la preposición ‘por’ o la preposición ‘de’, por lo que se considera a la nominalización ‘interpretación’ subespecificada, y no resultativa como ADN proponía.

El restante 5 % de error se explica porque ADN es incapaz de detectar un patrón que responda a las nominalizaciones subespecificadas como por ejemplo son el hecho de que derive de un verbo que exprese un logro y que tenga un arg1-tem (tema). En el ejemplo (12), la clasificación se clasifica como subespecificada y no como resultativa, como lo hacía ADN, porque el verbo base de ‘desenlace’ es un logro y porque aparece el argumento tema (arg1-tem).

Finalmente, los casos en los que ADN clasificó nominalizaciones subespecificadas como construcciones lexicalizadas se explican porque el clasificador confundió estas ocurrencias con construcciones lexicalizadas que compartían el mismo lema (el porcentaje de error es del 1 %). En el ejemplo (13) la confusión se produce con la lexía ‘pensión de jubilación’.

Lexicalizaciones

Las construcciones lexicalizadas que fueron incorrectamente clasificadas por ADN en la mayoría de los casos (96 %, 90 instancias) fueron clasificadas como resultativas. Esto se explica muy probablemente porque las construcciones lexicalizadas nominales, a las que sí se les asigna tipo denotativo, son en su mayoría de la clase resultativa. Por lo tanto, ADN falla principalmente en distinguir entre los diferentes tipos de construcciones lexicalizadas.

(14) Defendió sin [**reservas**<lexía no nominal>]SN a su compañero.

En el ejemplo anterior, ADN no ha distinguido que esta lexicalización, ‘sin reservas’, no es nominal, por lo que no es un resultado (denotación típica de las lexías nominales) sino adverbial y, por lo tanto, no tiene clase denotativa específica.

Nos hemos basado en el modelo LEAFF que daba una tasa de error de 12,8 % (ver la Tabla 7.1). La distribución de estos errores entre los tres tipos denotativos y las lexías es la siguiente: 1,66 % para los resultativos, 5,63 % para los eventivos, 5,50 % para los subespecificados y 0,01 % para las lexías.

Los porcentajes de errores que atribuimos a un mal etiquetado manual se deben entender como cifras relativas al porcentaje de errores de ADN. Es decir, el 27 % de errores manuales para el caso de los resultativos corresponde en términos absolutos a un 0,44 % (1,66 x 0,27); el 51 % de errores manuales en los eventivos se corresponde a un 2,87 % (5,63x 0,51); y, el 45 % de los errores manuales en los subespecificados se corresponde a un 2,47 % (5,50x0,45). Todas estas cantidades están dentro de los márgenes de acuerdo entre anotadores que aparecen en la Tabla 8.2 del Capítulo 8.

7.4. Discusión

Como vimos en el Capítulo 2, existen algunos trabajos que se centran en el tratamiento computacional de las nominalizaciones deverbales pero están básicamente interesados en 1) la detección de relaciones semánticas entre sustantivos, Tarea 4 de SemEval 2007 (Girju et al., 2009) y Tarea 8 de SemEval 2010 (Hendrickx et al., 2010) o en compuestos nominales Girju et al. (2005) y Tarea 9 de SemEval 2010 (Butnariu et al., 2009, 2010) por un lado; y 2) la asignación de roles semánticos a las nominalizaciones a partir de información verbal (Hull and Gomez, 2000; Lapata, 2002; Padó et al., 2008; Gurevich and Waterman, 2009), por el otro. A pesar de que la mayoría de estos trabajos manifiestan un conocimiento de la distinción lingüística de evento y resultado, ninguno de ellos aplica esta distinción en sus sistemas. La noción de evento aparece en el trabajo de Creswell et al. (2006), cuyo clasificador distingue entre ocurrencias nominales que denotan eventos y no-eventos. Sin embargo, en su trabajo no solo tratan con nominalizaciones sino con todo tipo de sustantivos por lo que su distinción no es comparable con la nuestra, como se vio en el Capítulo 2.

Creswell et al.,2006

De hecho, el único trabajo que se relaciona más estrechamente con el nuestro es el de Eberle et al. (2011), que trabajan con las nominalizaciones en *-ung* del alemán. En su trabajo se establece que estas nominalizaciones pueden denotar o un evento, o bien un estado o un objeto y restringen su muestra de nominalizaciones en *-ung* a aquellas que derivan de verbos de dicción ('decir', 'explicar', 'comunicar', etc.) y que se encuentran dentro de SPs introducidos por la preposición *nach* ('hacia') del alemán. Según los autores, este tipo concreto de nominalización puede denotar o bien un evento o una proposición, que es un tipo de objeto. Disponen de una herramienta de análisis semántico Eberle et al. (2008) que desambigua el tipo de nominalización teniendo en cuenta nueve criterios a los que ellos llaman indicadores. La herramienta proporciona una representación semántica a partir de la cual se extraen los indicadores y se calcula el tipo de nominalización a partir del esquema de peso predeterminado. Los autores aplican esta herramienta a 100 oraciones en las que los nueve indicadores están presentes y la herramienta reconoce correctamente el tipo de nominalización en un 82 % de los casos.

Eberle et al.,2011

Dado que ADN se sustenta en técnicas de ML y no restringe las nominalizaciones a un sufijo específico ni a un tipo específico de verbo base, el trabajo de Eberle et al. (2011) no es directamente comparable con el nuestro. Sin embargo, para solventar esta cuestión hemos replicado los experimentos seleccionando solo aquellas nominalizaciones creadas con el sufijo *-ción* (el sufijo más productivo del español y el más cercano al sufijo *-ung* del alemán) y derivadas de un verbo de dicción. Este subconjunto incluye 66 lemas de nominalizaciones de las 1.655 con las que trabajamos normalmente. Hemos aplicado el modelo LEAFF a las 719 ocurrencias del corpus correspondientes a estas 66 nominalizaciones y hemos ob-

tenido un 85,6 % de corrección. Esto supone una mejora del 3,6 % respecto a su resultado, a pesar de que nuestro modelo no está entrenado con este tipo concreto de nominalizaciones y de que no tiene criterios específicos para ellas. Aunque tenemos que tomar este resultado con mucha precaución dada la distancia entre las dos lenguas que comparamos, podemos decir que ADN consigue un buen resultado.

7.5. Conclusiones

En este capítulo se han detallado los experimentos realizados para desarrollar los modelos de clasificación (a nivel de sentido y a nivel de lema) a partir de los recursos AnCora-Nom-v3 y AnCora-Es-v3, unos modelos que aprenden con un mayor número de instancias y con recursos totalmente validados. Además también se han replicado los experimentos a nivel de sentido y lema con el subconjunto de AnCora-Es-v3 de 100.000 palabras y el subconjunto de 817 entradas léxicas de AnCora-Nom-v3. En total han resultado 32 modelos que responden a cinco dimensiones distintas. La conclusión más importante es que en modelos a nivel de lema, modelos que responden a una tarea de clasificación más realista, necesitan atributos tanto del léxico como del corpus para conseguir un buen resultado. Además, entre los 32 modelos generados se ha seleccionado el modelo más adecuado para cada escenario, ejemplificando, por tanto, como actuaría ADN en cada uno de ellos.

Parte IV

Recursos

CAPÍTULO 8

ANCORA-ES: VALIDACIÓN MANUAL

En este capítulo presentamos los dos procesos de validación manual que se llevaron a cabo para garantizar la calidad de la anotación de las nominalizaciones deverbales en el corpus AnCora-Es. El primero de ellos atañe a la estructura argumental de las nominalizaciones deverbales (Sección 8.1) y el segundo a la denotación de las mismas (Sección 8.2). Cada uno de estos procesos manuales se realizó en momentos diferentes de la investigación pero los agrupamos en el mismo capítulo porque la metodología seguida es la misma: ambos procesos se han apoyado en guías de anotación convenientemente seguidas por los anotadores y en pruebas de acuerdo entre los anotadores para garantizar la consistencia de la validación manual. En primer lugar, nos centramos en la validación manual de la estructura argumental que fue la primera que se realizó (Sección 8.1) y en segundo lugar abordaremos la validación manual de la denotación de las nominalizaciones (Sección 8.2). A continuación, presentamos la herramienta AnCora-Pipe y su adaptación para la validación manual de estos dos tipos de información (Sección 8.3). Finalmente veremos en las conclusiones, el resultado de este proceso de anotación, el corpus final AnCora-Es-v3 (Sección 8.4).

8.1. Validación manual de la estructura argumental

Como se ha visto en el Capítulo 4, la anotación de la estructura argumental del corpus AnCora-Es (500.000 palabras) se llevó a cabo de manera automática a partir del paquete de reglas heurísticas RHN, allí explicadas. La validación manual del corpus tiene una triple función: 1) sirve para garantizar la calidad de la anotación y la coherencia y consistencia de los datos anotados; 2) se ha utilizado como *gold standard* para evaluar el proceso automático, es decir, para evaluar la fiabili-

dad de las reglas heurísticas aplicadas (RHN) y evaluar las hipótesis lingüísticas que subyacen en dichas reglas; 3) con esta validación manual se obtiene el corpus AnCora-Es-v2, a partir del cual se realiza la inducción automática del léxico nominal AnCora-Nom-v2, ambos recursos utilizados por la segunda versión del clasificador ADN (ADN-Classifer-v2) para la anotación automática de la denotación en el corpus.

Anotadores manuales

Guía de anotación

Para conseguir el primer objetivo, esto es, garantizar la calidad de la anotación, se seleccionaron tres anotadores graduados en lingüística con experiencia en la anotación de la estructura argumental de los verbos del mismo corpus. Los anotadores disponían de una guía de anotación (Peris, 2011) en la que se describen los criterios lingüísticos y que incluye el esquema de anotación, la manera de proceder en la anotación y en la que se proporcionan ejemplos de anotación. En esta guía se pone de manifiesto la hipótesis de que las nominalizaciones deverbales heredan su estructura argumental del verbo base; de hecho, en el léxico verbal AnCora-Verb se consultan los argumentos y papeles temáticos asociados al verbo base que son a su vez los que se pueden asociar a la nominalización.

AnCora-Pipe

Para llevar a cabo la validación manual, se ha utilizado la herramienta de anotación AnCora-Pipe (Bertran et al., 2008) para minimizar los posibles errores (por ejemplo, solo se pueden utilizar las combinaciones de argumentos y papeles temáticos admitidas en el esquema de anotación e impide la asociación de argumentos a constituyentes no argumentales) y facilitar la tarea a los anotadores reduciendo el tiempo de anotación. Este proceso de validación manual fue precedido de una prueba de acuerdo entre anotadores con el fin de verificar que los anotadores habían entendido correctamente los criterios de anotación y cómo proceder y, por lo tanto, garantizar la consistencia y coherencia de la anotación final.

A continuación describimos en qué consiste específicamente este proceso de validación manual (Subsección 8.1.1). En segundo lugar, se presentan los criterios de anotación (Subsección 8.1.2). Finalmente, nos centramos en las pruebas de acuerdo entre anotadores (Subsección 8.1.3). En la Figura 8.1 se resume el proceso de validación manual de la estructura argumental en el corpus.

8.1.1. Descripción de la tarea de validación manual

La validación manual (Peris et al., 2010b) tiene como objetivo comprobar que la asignación automática de argumentos (arg0, arg1, arg2, etc.) y sus correspondientes papeles temáticos (agente, paciente, tema, etc.) es correcta. Es decir, que este proceso de validación manual se ha centrado en las nominalizaciones identificadas automáticamente, esto es, las ocurrencias en el corpus AnCora-Es de las 1.655 nominalizaciones seleccionadas manualmente (Sección 4.1). El total de nominalizaciones revisadas es de 24.864. Concretamente, los anotadores tenían que

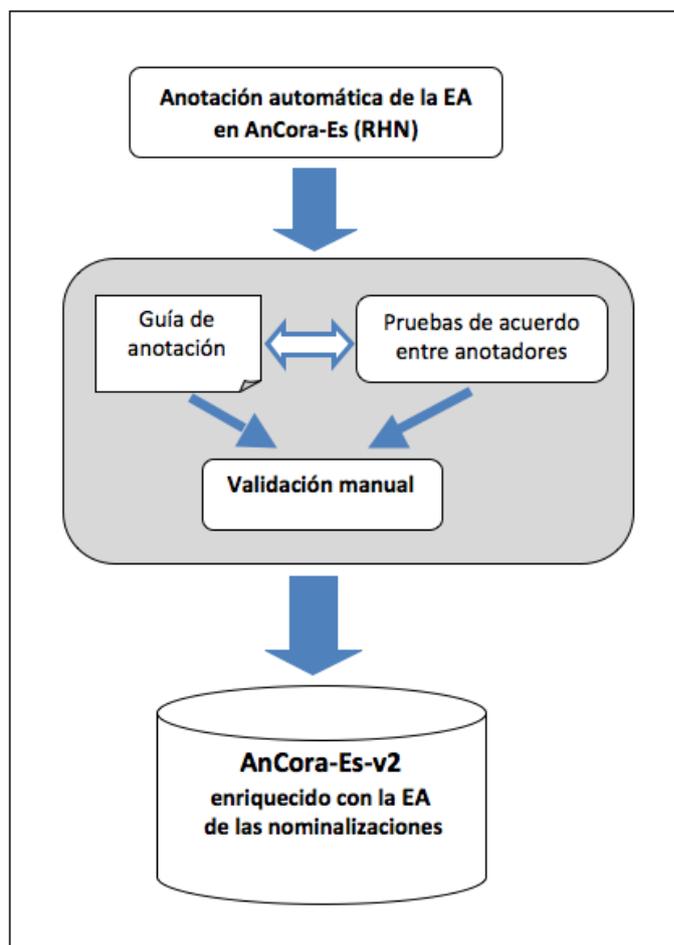


Figura 8.1: Validación manual de la estructura argumental de las nominalizaciones deverbales en AnCora-Es

validar tres tipos de información durante este proceso:

1. Debían cerciorarse que el sustantivo constituía realmente una nominalización deverbal, esto es, que exhibía propiedades verbales. Por ejemplo, el sustantivo ‘establecimiento’ tiene dos sentidos, puede interpretarse como sinónimo de ‘tienda’ o bien como ‘el proceso o resultado de establecer’. Dado que todas las formas de ‘establecimiento’ se anotaron automáticamente independientemente de su significado, era necesario que los revisores se aseguraran que el sustantivo anotado era realmente una nominalización deverbal.

Propiedades
deverbales

Sentido verbal

2. Los anotadores también tenían que revisar el sentido y el esquema sintáctico -semántico (*frame*) del verbo base, es decir, la información asociada al atributo <originlexicalid> donde se especifica si se trata de un esquema sintáctico-semántico transitivo, pasivo, inacusativo, etc., con el que se había asociado la nominalización deverbal en el proceso automático. Recuérdese que por defecto se elegía el sentido y el correspondiente esquema sintáctico-semántico verbal con un mayor número de argumentos. Los anotadores debían comprobar si dicho esquema sintáctico-semántico verbal era correcto, puesto que en función del esquema sintáctico-semántico verbal seleccionado, los argumentos disponibles para ser asociados a las nominalizaciones pueden variar.

Argumentos y papeles temáticos

3. Finalmente, los anotadores debían validar los argumentos (es decir, el atributo <arg>) y papeles temáticos (es decir, el atributo <tem>) y la etiqueta RefMod (complementos no argumentales) asociados automáticamente a los constituyentes de los SNs deverbales. Además, si consideraban que un sustantivo tenía un argumento incorporado anotaban el sustantivo con el argumento y papel temático correspondiente. Esto último se hace con el mismo esquema de anotación que se ha descrito en el Capítulo 3 (Sección 3.2).

8.1.2. Criterios de anotación

Para llevar a cabo la tarea de validación de los argumentos internos anotados automáticamente y para anotar manualmente los argumentos incorporados, los anotadores seguían los siguientes criterios:

Esquema de anotación

1. Utilizar el mismo esquema de anotación empleado en la anotación automática, es decir, seis posibles valores de posición argumental combinados con los 19 papeles temáticos posibles y la etiqueta no argumental RefMod para aquellos complementos del nombre no argumentales (Véase Sección 3.2).

Constituyentes

2. Tener en cuenta el tipo de constituyente al que está asociado el argumento y papel temático. En principio, automáticamente solo debían anotarse como argumentos de los sustantivos, los SPs, los SA que tienen como núcleos adjetivos relacionales, los Grel y los Poss. El resto de constituyentes, es decir, los SAs de núcleo no relacional, los SAdv, los SNs o las Ss, se les asignaba la etiqueta RefMod (Sección 3.3). Aunque en general, se había visto que en español el primer grupo de constituyentes era mayoritariamente argumental y el segundo no, ya en el estudio empírico (Capítulo 3) se observó que podían existir SPs, por ejemplo, no argumentales (1), y SAdv que podían interpretarse como argumentos adjuntos (2). Por lo tanto, los

revisores debían verificar si los constituyentes del primer grupo eran realmente argumentales y si los del segundo grupo no lo eran.

- (1) Él es el jugador del equipo que mejor porcentajes tiene en [**tiros** [de tres]_{SP-RefMod}]_{SN}.
- (2) Hizo [**declaraciones** a los periodistas [antes de visitar la feria de arte Arco]_{SAdv-argM-tmp}]_{SN}.

3. Consultar las entradas verbales en el léxico AnCora-Verb. Esta consulta era sumamente importante y la base del proceso de validación manual. Por una parte, servía para validar o no el sentido verbal y su correspondiente esquema sintáctico-semántico (*frame*). Por otra parte, y en función del sentido verbal seleccionado, los anotadores sabían qué argumentos podían ser candidatos a ser argumentos de la nominalización de la que aquel verbo se derivaba. Veámoslo con detenimiento.

Consulta de
AnCora-Verb

- a) Validación del sentido verbal. La elección del sentido verbal del que deriva la nominalización no es una decisión trivial. La asignación de los argumentos y sus correspondientes papeles temáticos a la nominalización viene determinada por los argumentos y los papeles temáticos asociados con el verbo base de dicha nominalización. Los sentidos verbales se caracterizan por tener estructuras argumentales distintas, por lo tanto, la correcta elección del sentido verbal es clave para determinar los argumentos de la nominalización. Recuérdese que automáticamente se asociaba el sentido verbal con un mayor número de argumentos en el caso de que hubiera polisemia. Sin embargo, este sentido no siempre era el adecuado. Por ejemplo, en (3), la nominalización ‘voladura’ estaba asociada al sentido del verbo ‘volar’ que significa ‘gravitar por el aire’ y no con el de ‘hacer saltar algo con violencia’ ya que el primero tenía tres argumentos nucleares (no adjuntos) y el segundo solo dos, pero esta asociación era incorrecta (Véase la Figura 8.2). El hecho de detectar el sentido correcto ha permitido que los argumentos asociados a la nominalización sean los correctos.

Sentido Verbal

- (3) El Foro por la Memoria pide [la **voladura** [de la cruz del Valle de los Caídos]_{SP-arg1-pat}]_{SN}.

- b) Validación de la asignación de la posición argumental y papel temático a los constituyentes internos al SN y anotación de los argumentos incorporados a la nominalización.

Argumentos y
papeles temáticos

En primer lugar, nos centramos en la validación de la asignación de la posición argumental y papel temático a los constituyentes internos al SN. Si uno de los constituyentes del SN puede ser interpretado como uno de los argumentos declarados en la entrada verbal, entonces se debe anotar el constituyente nominal con dicha posición argumental y papel temático. Por ejemplo, en el caso de (3) el sentido de ‘volar’ con el significado de ‘hacer saltar algo con violencia’ tiene dos argumentos, un paciente (arg1-pat) y un agente (arg0-agt) (Véase la Figura 8.2); el anotador tiene que validar si el complemento SP ‘de la cruz del Valle de los Caídos’ recibe la interpretación semántica correcta, es decir, si se trata efectivamente de un arg1-pat.

type	verb	lemma	volar
lng	es	origin	NOT PRESENT

- id: 1[verb.volar.1]
 - default [verb.volar.1.default] – D11.inergative-agentive : [suj/arg0/agt]+[cc/arg3/ori]+[cc/arg4/des]+[argM/(loc)]
 - suj/arg0/agt
 - cc/argM/loc
 - cc/arg3/ori
 - cc/arg4/des
 - EXAMPLES (3)
 - ▶ volar alrededor del mundo
 - ▶ la turbina del avión de Spanair que volaba de Valencia a Dublín
 - ▶ el tiempo
- id: 2[verb.volar.2]
 - default [verb.volar.2.default] – A21.transitive-agentive-patient : [suj/arg0/agt]+[cd/arg1/pat]
 - suj/arg0/agt
 - cd/arg1/pat
 - EXAMPLES (1)
 - ▶ que pretendía volar un furgón ocupado por 24 militares , según la justicia española .
- id: 3[verb.volar.3]
 - default [verb.volar.3.default] – B21.unaccusative-state : [suj/arg1/tem]
 - suj/arg1/tem
 - EXAMPLES (1)
 - ▶ el tiempo vuela a la velocidad de los cambios tecnológicos

Figura 8.2: Entrada léxica del verbo ‘volar’ en AnCora-Verb

Si no puede ser interpretado como ninguno de los argumentos de la entrada verbal, entonces existen dos posibilidades: (a) el constituyente se interpreta como un argM que no está representado en la entrada verbal (4) o (b) el constituyente no tiene una interpretación argumental pero es un modificador de la nominalización, por lo que se le asigna la etiqueta RefMod (1).

- (4) El Foro por la Memoria pide [la **voladura** [de la cruz del Valle de los Caídos]SP-arg1-pat [la semana que viene]SN-argM-tmp]SN.

En segundo lugar, prestamos atención a la anotación de los argumentos incorporados a la nominalización. Como hemos visto en el apartado (Sección 3.3), en algunas ocasiones los argumentos de la nominalización se encuentran incorporados en el propio sustantivo (5), esto es, el sustantivo puede denotar un argumento del verbo base. Si tras consultar AnCora-Verb, alguno de los argumentos declarados en la entrada léxica verbal encaja en la interpretación del sustantivo, entonces le asignamos dicho valor argumental.

Argumentos
incorporados

En el caso de los argumentos incorporados los valores de <arg> y <tem> se declararán en el nodo nombre. El hecho que dichos atributos se marquen a nivel de nombre nos indica que los argumentos están incorporados al sustantivo. Esto nos sirve para diferenciarlos del resto de argumentos nominales que se marcan a nivel de constituyente. También se debe tener en cuenta que aunque reciban argumento y papel temático, estos sustantivos nunca reciben la etiqueta CN porque no son complementos de sí mismos. En el ejemplo (5), se puede observar que el arg1-pat está incorporado al sustantivo ‘propuesta’.

- (5) IC-V ha planteado [una **propuesta**_{arg1-pat} [a Joan Clos]_{SP-arg2-ben}]_{SN}.

A partir de estos criterios se ha llevado a cabo la validación manual de la estructura argumental de las nominalizaciones deverbales del corpus AnCora-Es. A continuación presentamos las pruebas de acuerdo entre anotadores que realizamos para garantizar la fiabilidad y la coherencia del proceso manual de anotación.

8.1.3. Pruebas de acuerdo entre anotadores

Las pruebas de acuerdo entre anotadores se llevaron a cabo previamente al proceso de validación manual. Sirvieron para comprobar que los anotadores habían entendido los criterios de anotación y la tarea que debían realizar. Además, sirvió para que los anotadores se familiarizaran con la herramienta AnCora-Pipe y la adaptación que de ella habíamos realizado para la anotación de las nominalizaciones deverbales (Sección 8.3). Era importante observar el grado de acuerdo entre los anotadores en la anotación de la estructura argumental de las nominalizaciones para garantizar la consistencia de los datos anotados y la calidad del proceso de validación.

Para llevar a cabo la prueba como muestra de datos seleccionamos de forma aleatoria 100 oraciones del corpus AnCora-Es que reunieran el requisito de

Muestra

a ser argumento de la nominalización; en total se incluyeron 131 constituyentes candidatos.

Anotadores

Participaron en estas pruebas tres estudiantes del grado de Lingüística de la Universidad de Barcelona que tenían experiencia previa en la anotación de la estructura argumental de los verbos del corpus AnCora-Es, por lo que no fue necesario ningún proceso de entrenamiento previo de los anotadores.

Tarea

En la prueba los anotadores tenían que decidir para cada constituyente, (a) si era un argumento, y en tal caso, (b) qué argumento y papel temático (de las 36 combinaciones posibles) le correspondía. Para ello, tuvieron que elegir el sentido verbal del que procedía la nominalización y tener en cuenta la información especificada en el léxico AnCora-Verb acerca de dicho sentido verbal. Los tres anotadores tenían que realizar esta tarea en paralelo, sin posibilidad de compartir información entre ellos.

Medidas de evaluación

La elección del sentido verbal es importante ya que se ha calculado el grado de acuerdo teniendo en cuenta si los anotadores estaban de acuerdo en el sentido verbal correspondiente a la nominalización. Esperamos un grado de desacuerdo alto cuando el sentido verbal elegido por los diferentes anotadores no es el mismo. Hemos calculado el grado de acuerdo usando el acuerdo observado—*observed agreement*— (Scott, 1955) y el coeficiente Kappa (Siegel and Castellan, 1988). El acuerdo observado mide simplemente la proporción de constituyentes en los que hay acuerdo respecto al total de la anotación.

$$\text{Acuerdo observado}(A_o) = \frac{\text{número de constituyentes en los que hay acuerdo}}{\text{total de constituyentes anotados}}$$

El coeficiente Kappa descuenta de esta proporción (A_o) la parte de acuerdo por azar. La medida Kappa es pues siempre inferior al acuerdo observado.¹

$$Kappa = \frac{(A_o - A_e)}{(1 - A_e)}$$

Sin embargo, hemos aumentado la penalización del grado de acuerdo en estas dos medidas si el sentido verbal era compartido entre los distintos anotadores y la hemos disminuido si el sentido verbal elegido no era el mismo. El esquema de peso² asignado es de 40 % en el primer caso y 60 % en el segundo caso.

$$\text{Acuerdo total} = (0,4 * \text{mismo sentido verbal}) + (0,6 * \text{diferente sentido verbal})$$

Resultados

En la Tabla 8.1 presentamos los resultados de la prueba de acuerdo entre anotadores. Las columnas muestran los resultados para cada pareja de anotadores y el resultado medio entre las tres parejas. Las filas presentan los resultados del acuerdo observado y del coeficiente Kappa de acuerdo con las fórmulas arriba mencionadas.

¹En esta fórmula A_e significa Acuerdo Esperado.

²El esquema de pesado fue definido empíricamente.

Parejas de anotadores	A y B	A y C	B y C	Resultado Global
Mismo Sentido Verbal	119	125	125	
Acuerdo Observado	86 %	96 %	90 %	90,6 %
Kappa	84 %	94 %	88 %	88,6 %
Diferente Sentido Verbal	12	6	6	
Acuerdo Observado	66 %	66 %	83 %	71,6 %
Kappa	60 %	58 %	80 %	66 %
Total	131	131	131	
Acuerdo Observado	74 %	78 %	85,8 %	79,2 %
Kappa	69,6 %	72,4 %	83,2 %	75 %

Tabla 8.1: Resultados de la prueba de acuerdo entre anotadores: estructura argumental

Nos centramos en el resultado global, la media entre las tres parejas. Como era de esperar, cuando los anotadores no estaban de acuerdo con el sentido verbal correspondiente a la nominalización, el acuerdo disminuye aproximadamente un 20 % tanto en el acuerdo observado (71,6 %) como en el kappa (66 %) respecto a cuando los anotadores sí están de acuerdo en el sentido verbal (90,6 % y 88,6 %, respectivamente). Como se ha mencionado anteriormente, es muy difícil lograr un alto grado de acuerdo si el sentido verbal elegido por los anotadores es diferente puesto que los argumentos y papeles temáticos para ser mapeados varían. Según la fórmula presentada, le media de acuerdo entre anotadores alcanza un 75 % de kappa, que se trasluce en un acuerdo observado del 79,2 %. Este es un nivel de acuerdo muy satisfactorio (Fleiss, 1981) teniendo en cuenta que contamos con 36 etiquetas semánticas, lo que supone más oportunidades para el desacuerdo. Por lo tanto, este nivel de acuerdo nos garantiza que el proceso de validación manual es fiable y coherente.

8.2. Validación manual de la denotación

La anotación automática de la denotación se llevó a cabo mediante una versión intermedia del clasificador ADN (ADN-Classifier-v2), que distingue entre nominalizaciones eventivas, resultativas y subespecificadas, además de reconocer qué nominalizaciones forman parte de construcciones lexicalizadas y de qué tipo son. En esta versión se adaptó a nivel de lema el modelo aprendido para sentidos (ADN-v1) en los experimentos iniciales realizados con el clasificador (Sección 5.2.1).

Como en el caso de la estructura argumental, la validación manual del corpus tiene una triple función: 1) sirve para garantizar la calidad de la anotación y, la coherencia y consistencia de los datos anotados; 2) se ha utilizado como *gold standard* para aprender y evaluar los diferentes modelos del clasificador ADN-v3; 3) esta validación manual, además, da lugar al corpus AnCora-Es-v3, a partir del cual se realiza la inducción automática del léxico nominal definitivo AnCora-Nom-v3, que describimos en el siguiente capítulo y que también ha sido usado como recurso para desarrollar la versión definitiva del clasificador, ADN-Classifier-v3.

Anotadores manuales

Con el objetivo de garantizar la calidad de la anotación, se seleccionaron los tres anotadores manuales que tras las pruebas de acuerdo, en las que participaron cinco personas, habían conseguido un mayor grado de acuerdo. Estas pruebas de acuerdo se realizaron previamente al proceso de validación manual para asegurar que los anotadores habían entendido correctamente la tarea y los criterios de anotación, y por lo tanto, garantizar la consistencia y coherencia de la anotación. También en este caso, se ha utilizado la herramienta de anotación AnCora-Pipe (Bertran et al., 2008) para minimizar los posibles errores y facilitar la tarea a los anotadores reduciendo el tiempo de validación.

AnCora-Pipe

A continuación, describimos en primer lugar en qué consiste específicamente este proceso de validación manual (Subsección 8.2.1). En segundo lugar, se presentan los criterios de anotación (Subsección 8.2.2). Finalmente, nos centramos en las pruebas de acuerdo entre anotadores (Subsección 8.2.3). En la Figura 8.3 se resume el proceso de validación manual de la denotación en el corpus.

8.2.1. Descripción de la tarea de validación manual

La validación manual (Peris et al., 2010b) tiene como objetivo comprobar los dos tipos de información que automáticamente asigna el clasificador ADN. Por una parte, se ha de verificar si el tipo denotativo (evento, resultado, subespecificado) asociado a las nominalizaciones es el correcto y, por otra, si las predicciones de que dichas nominalizaciones forman parte de construcciones lexicalizadas o no son correctas. Además, los anotadores debían verificar el tipo de lexicalización (nominal, verbal, adjetival, preposicional, adverbial o conjuntiva) en el caso de que se hubiera considerado que la nominalización formaba parte de una construcción lexicalizada. Este proceso de validación manual se ha realizado sobre las 23.431 ocurrencias que se habían verificado como realmente deverbales en el anterior proceso de validación manual.

Para llevar a cabo este proceso de validación manual se proporcionó a los anotadores una serie de criterios lingüísticos para poder distinguir cuando una nominalización formaba parte de una construcción lexicalizada y para clasificar las nominalizaciones según su denotación. Dado que la distinción con la que trabajábamos en este proceso de validación no es fácil y que los anotadores no esta-

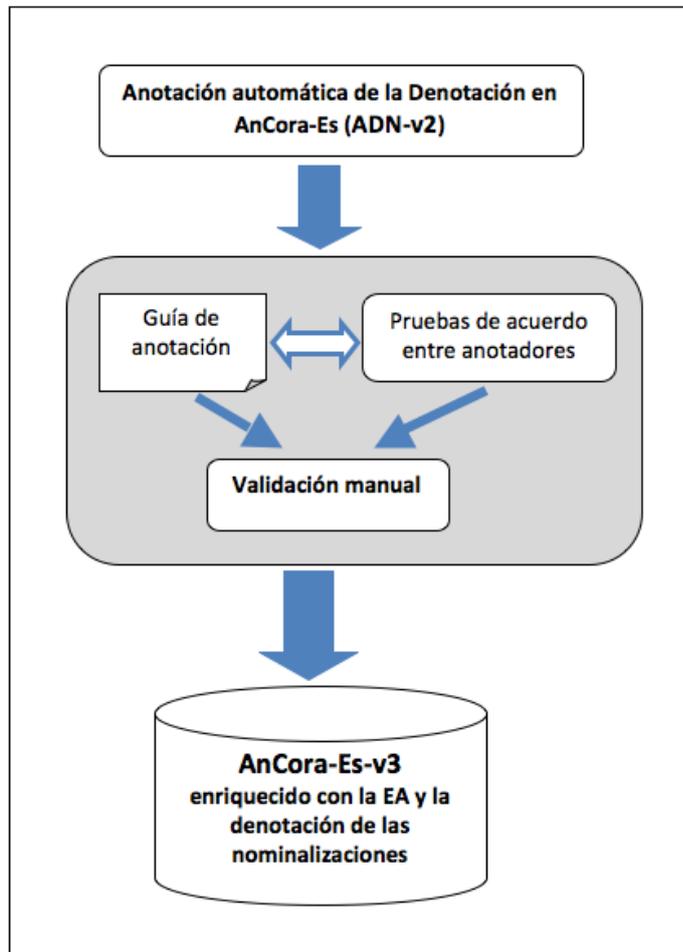


Figura 8.3: Validación manual de la denotación de las nominalizaciones deverbales en AnCora-Es

ban familiarizados con los criterios, quisimos en primer lugar realizar unas pruebas de acuerdo entre anotadores. De los cinco anotadores que participaron en las pruebas de acuerdo, solo los tres con mayor índice de acuerdo y que, por lo tanto, demostraron haber entendido mejor los criterios, participaron en el proceso final de validación manual.

8.2.2. Criterios lingüísticos para la clasificación de las nominalizaciones de verbales según su denotación

En esta subsección detallamos los criterios morfológicos, sintácticos y semánticos consignados a los anotadores para la clasificación de las nominalizaciones de verbales en eventivas, resultativas y subespecificadas. La mayoría de estos criterios fueron definidos durante el proceso de análisis empírico de la denotación (Capítulo 5), es decir, son criterios tomados de la bibliografía confirmados como útiles para el español (Subsección 5.1.1) o criterios que emergieron de ese proceso, tanto durante el análisis lingüístico (Subsección 5.1.2) como computacional (Subsección 5.2.2).

Lexicalizaciones

Sin embargo, antes de aplicar estos criterios los anotadores tenían que decidir si la nominalización formaba parte de una construcción lexicalizada. Y en tal caso, validar o asignar el tipo de lexicalización: nominal, verbal, adjetival, preposicional, adverbial y conjuntiva. Recuérdese que solo a las lexicalizaciones nominales se les asigna tipo denotativo. En (6), ‘golpe de estado’ se considera una construcción lexicalizada por tres razones. En primer lugar, la construcción en global tiene una referencia distinta a la que tiene la nominalización por sí sola, esto es, un ‘golpe de estado’ no es un tipo de golpe sino una actuación, normalmente militar, de signo político. En segundo lugar, el segundo elemento de la nominalización, ‘estado’, no puede tomar complementos por cuenta propia, como sería en este caso ‘democrático’ (7). Finalmente, si insertamos un elemento en la construcción lexicalizada el resultado no es gramatical (8).

(6) Se habla de [un **golpe de Estado**<lexicalización-resultado>]_{SN} de manera irresponsable.

(7) *Se habla de [un **golpe de Estado** democrático]_{SN} de manera irresponsable.

(8) *Se habla de [un golpe de gran Estado]_{SN} de manera irresponsable.

Tipo denotativo

Una vez que los anotadores deciden si la nominalización es parte de una construcción lexicalizada (y en tal caso, el tipo de lexicalización), deben comprobar que el tipo denotativo –evento, resultado, subespecificado– sea correcto en función de los criterios que presentamos a continuación. Estos criterios no son determinantes individualmente, sino que deben ser entendidos como indicadores, la combinación de los cuales les ayuda a decidir el tipo denotativo adecuado para cada nominalización.

■ Criterio de la paráfrasis

Fruto del análisis lingüístico, se estableció como un primer criterio holístico y general, la posibilidad de parafrasear la nominalización por una estructu-

ra clausal como indicador de que dicha nominalización es un evento y la imposibilidad como indicador de la lectura resultativa (Veáse la Subsección 5.1.2).

■ **Argumento Incorporado**

En el estudio empírico se observó que en español, una nominalización es resultativa si incorpora el argumento interno (arg1) del verbo base correspondiente. Por ejemplo, ‘invento’ denota un objeto resultante del verbo base ‘inventar’, esto es, la nominalización tanto se refiere a la acción verbal de inventar como al objeto resultante del verbo (9). Esta última lectura queda reforzada por el hecho de que es imposible encontrar un constituyente en el SN deverbal que reciba la interpretación de argumento incorporado, es decir que se realice como argumento paciente.

(9) [El **invento**_{arg1-pat} <resultado> de Juan]_{SN} tuvo mucho éxito.

■ **Pluralidad**

Uno de los criterios propuestos en la bibliografía (y confirmado en el estudio empírico) para identificar las nominalizaciones resultativas es su posibilidad de aparecer en plural (10), contrariamente a las nominalizaciones eventivas (11).

(10) Para compensar [las **pérdidas**_{<resultado>} ante sus depredadores]_{SN}, los tíes traen al mundo gemelos.

(11) [...] aunque [la **pérdida**_{<evento>} del pivot Rodney Dent]_{SN} puede condenar a los de Rick Pitino.

■ **Determinantes**

Otro de los criterios comúnmente aceptados en la bibliografía y verificado como útil en el estudio empírico para la distinción entre evento y resultado en las nominalizaciones deverbales es el tipo de determinante que estas aceptan. Las nominalizaciones eventivas pueden ser especificadas por un artículo definido, un determinante posesivo o bien pueden aparecer sin especificación alguna (12). Las nominalizaciones resultativas pueden además estar especificadas por determinantes demostrativos, artículos indefinidos y numerales (13).

(12) No fue un hecho aislado, sino [la **culminación**_{<evento>} de [una dinámica de deterioro y **deslegitimación**_{<evento>} de las instituciones por parte del PP]_{SN}]_{SN}.

- (13) Las exportaciones totales pasaron de los 12,3 millones de dólares en 1999 a los 14,8 millones en el presente año, lo que supone [una **subida**<resultado> del 20,47 por ciento]SN.

■ **Complementación**

De la bibliografía consultada dos criterios referentes al tipo de complementación se confirmaron como relevantes para caracterizar a las nominalizaciones según su tipo denotativo. Por una parte, los adjetivos relacionales se ratificaron como argumentos de las nominalizaciones resultativas (14) pero no de las eventivas (15). De hecho, (15) es un ejemplo agramatical porque 'producción' no puede ser entendido como un evento: la interpretación de 'quesera' como arg1 bloquea la lectura eventiva.

- (14) El tema de conversación era [la **actuación**<resultado> [policial]SA-arg0-agt]SN.

- (15) *[La **producción**<evento> [quesera]SA-arg1-pat por los holandeses]SN.

Por otra parte, aunque tanto las nominalizaciones resultativas como las eventivas pueden aparecer con adjuntos temporales, los que complementan a las nominalizaciones resultativas deben estar introducidos por la preposición 'de' (16), mientras que dicha preposición no es necesaria en el caso de los que complementan a las nominalizaciones eventivas (17).

- (16) Hoy, tras [una **negociación**<resultado> [de trece horas]SP-argM-tmp]SN, se ha aprobado un nuevo texto sobre la reforma del seguro de desempleo.

- (17) La compañía presentó una auditoría por primera vez desde [su **constitución** <evento> [en 1989]SP-argM-tmp]SN

■ **Clase verbal**

Durante el análisis empírico se observó que la clase semántica del verbo base era de mucha utilidad para anotar la denotación. Las nominalizaciones se anotan teniendo en cuenta el sentido del verbo base (atributo <originlexicallid>) al que le corresponde una determinada clase semántica especificada en el léxico AnCora-Verb. Recuérdese que en este léxico existen 12 clases que pertenecen a 4 grandes grupos definidos de acuerdo con las clases aspectuales de Vendler (1967): realizaciones, logros, estados y actividades. Los sustantivos derivados de verbos de la clase semántica de las realizaciones y de los logros pueden dar lugar a nominalizaciones resultativas, eventivas y

subespecificadas. La lectura de la nominalización depende de qué argumentos se realizan en el SN y cuáles son los constituyentes que los explicitan, tal y como se ha resumido en la Tabla 5.9 y la Tabla 5.8 (Subsección 5.2.2). Los sustantivos derivados de verbos estativos y de actividad indican mayoritariamente una lectura resultativa como quedó establecido en la Subsección 5.1.1.

■ Selectores

Cuando los criterios anteriores no indican de manera clara el tipo denotativo, en el estudio empírico encontramos otros indicadores que pueden ser de ayuda para esclarecer la denotación, son los llamados selectores. Aunque en los trabajos de Balvet et al. (2010) se habla de criterios de este tipo para el francés y en Eberle et al. (2009) se habla de indicadores para el alemán, lo cierto es que este tipo de criterio es específico de cada lengua porque tiene que ver más con significados específicos de los lexemas (diferentes en cada lengua) que no con características morfosintácticas o semánticas. Distinguimos dos tipos de selectores: los selectores externos, es decir, los elementos que desde fuera del SN indican la denotación de la nominalización; y los selectores internos, es decir, prefijos de la nominalización que propician un tipo concreto de denotación (18). Como selectores externos incluimos preposiciones (19), sustantivos (20), adjetivos (21), verbos (22) y adverbios (23).

- (18) Hoy [la **reconstrucción**_{<evento>} de la ciudad]_{SN} llevará años.
- (19) Tras [la **presentación**_{<evento>} de este escrito]_{SN}, el titular de Fomento deberá comparecer ante la comisión competente del Senado.
- (20) De ahí su intento del [**cambio**_{<evento>} de fechas para la disputa de la próxima edición de la Vuelta]_{SN}.
- (21) Una de las primeras formas de piel tuvo que ser algo así como una membrana, resultante d[el **endurecimiento**_{<evento>} de la sustancia celular]_{SN}³.
- (22) El segundo proviene de [la **emisión**_{<evento>} de electrones rápidos]_{SN}.

³Este ejemplo es el mismo que el de la Subsección 5.1.2 porque no hemos encontrado un adjetivo distinto a ‘resultante’ que nos proporcione una pista sobre la denotación de las nominalizaciones.

(23) Una generación en vías de [**extinción**_{<evento>}]SN⁴.

■ **Criterio del agente**

En la Tabla 5.9 del Capítulo 5 se establece que cuando un argumento agente (arg0-agt) se realiza mediante un SP introducido por la preposición ‘por’ o ‘por parte de’, la nominalización resultará eventiva, mientras que si la preposición es ‘de’ o ‘entre’, la nominalización será resultativa. En el caso de que ambas preposiciones sean posibles, la nominalización será de tipo subespecificado. Dado que este argumento no es muy frecuente en el corpus y, sin embargo, es claramente desambiguador, se estableció que los anotadores podían inferir qué tipo de argumento agentivo sería más adecuado para la nominalización que estuvieran evaluando (Subsección 5.1.2). En el ejemplo (24), la denotación asociada a la nominalización es subespecificada porque tanto la preposición ‘de’ como la preposición ‘por’ parecen posibles para la realización del argumento agente (arg0-agt) y porque el contexto de la oración es insuficiente para establecer si se refiere a un evento o a un resultado.

(24) Anunció que el gabinete ha aprobado varias medidas económicas, como bajar un punto el IVA; continuar los esfuerzos para reducir la inflación; [la **aprobación**_{<subespecificado>} [del proyecto de ley de telecomunicaciones]SP-arg1-pat]SN.

(25) Anunció que el gabinete ha aprobado varias medidas económicas, como bajar un punto el IVA; continuar los esfuerzos para reducir la inflación; [la **aprobación**_{<resultado>} [del consejo de ministros]SP-arg0-agt [del proyecto de ley de telecomunicaciones]SP-arg1-pat]SN.

(26) Anunció que el gabinete ha aprobado varias medidas económicas, como bajar un punto el IVA; continuar los esfuerzos para reducir la inflación; [la **aprobación**_{<evento>} [por el consejo de ministros]SP-arg0-agt [del proyecto de ley de telecomunicaciones]SP-arg1-pat]SN.

La Figura 8.4 presenta la plantilla utilizada por los anotadores que resume los criterios de anotación descritos para la anotación de la denotación en las nominalizaciones deverbales⁵.

⁴Este ejemplo es el mismo que el de la Subsección 5.1.2 porque no hemos encontrado un adverbio o locución adverbial distinto a ‘en vías de’ que nos proporcione una pista sobre la denotación de las nominalizaciones.

⁵En la Figura 8.4 los nombres de los tipos denotativos que deben ser asociados por los anotadores manuales aparecen en inglés porque en la herramienta utilizada para la anotación, AnCoraPipe, los nombres de atributos y valores son en inglés. De esta manera, evitamos una posible confusión a los anotadores.

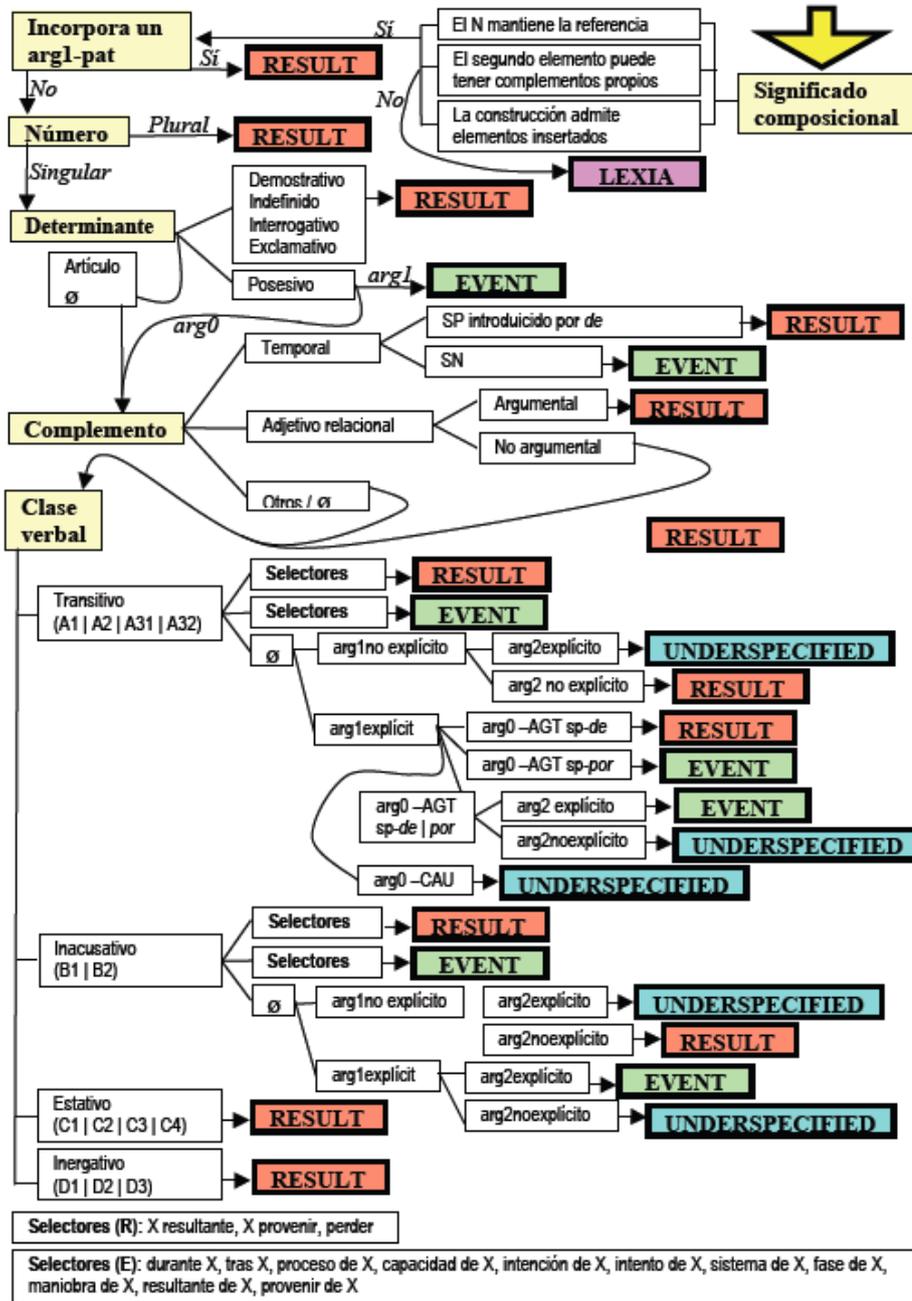


Figura 8.4: Aplicación de los criterios para la distinción Evento vs. Resultado

8.2.3. Pruebas de acuerdo entre anotadores

Como en el caso de la estructura argumental, se llevó a cabo una prueba de acuerdo entre anotadores para garantizar la fiabilidad y coherencia de la validación manual.

Anotadores

Cinco estudiantes del grado de Lingüística de la Universidad de Barcelona participaron en la prueba que tuvo dos etapas: primeramente, como ninguno de ellos tenía experiencia previa en la distinción denotativa que estamos tratando y esta no es una distinción semántica que resulte fácil, llevamos a cabo un proceso de entrenamiento que permitió la familiarización con los criterios y su aplicación, en el que, además, se discutieron los desacuerdos, lo que permitió comprobar la claridad de la guía de anotación. Finalmente, se realizó la prueba real a partir de la cual calculamos el nivel de acuerdo final.

Muestra de datos

La muestra de datos para el proceso de entrenamiento consistía en 100 oraciones que contenían cada una de ellas una nominalización seleccionadas aleatoriamente del corpus AnCora-Es. La muestra de datos para la prueba de acuerdo era de 200 oraciones en las que había una nominalización.

Medidas de evaluación

El nivel de acuerdo se calculó con las medidas de acuerdo observado y kappa descritas en la Sección 8.1.3.

Resultados

La Tabla 8.2 presenta los resultados globales de la prueba de acuerdo entre anotadores de la denotación.

Resultado de la media por parejas	Entrenamiento	Prueba
Acuerdo Observado	68 %	75 %
Kappa	44 %	60 %
Muestra de datos	100 oraciones	200 oraciones

Tabla 8.2: Resultados de la prueba de acuerdo entre anotadores: denotación

Como era de esperar, existe una mejora en la prueba final respecto al entrenamiento, que es incluso más remarcable en el coeficiente kappa (16 puntos de mejora). En cuanto a los resultados de la prueba real, se puede decir que el nivel de acuerdo entre anotadores es moderadamente bueno (60 % de Kappa, 75 % de acuerdo observado) teniendo en cuenta que la distinción semántica que tratamos no es nada fácil⁶. Además para garantizar aún más la calidad de la anotación, elegimos de los 5 anotadores aquellos (un total de 3) que habían logrado un acuerdo observado del 80 % (un kappa del 65 %). Estos tres anotadores podían consultarse

⁶Es comúnmente aceptado que un nivel de kappa por encima del 75 % es excelente, de un 40 % a un 75 % es de correcto a bueno, y por debajo del 40 % es pobre (Fleiss, 1981). Según esto, un resultado de 60 % de kappa es un buen nivel de acuerdo.

durante el proceso de validación manual, lo que no estuvo permitido en la prueba de acuerdo. Estos hechos han permitido asegurar la consistencia y la calidad de la anotación manual de la denotación en el corpus AnCora-Es y, a su vez, evaluar el funcionamiento del clasificador ADN.

8.3. Adaptación de AnCora-Pipe para la anotación de los SNs

AnCora-Pipe (Bertran et al., 2008) es un entorno informático para la creación, edición y análisis de corpora lingüísticos y lexicones. AnCora-Pipe está implementado como *plugin* en la plataforma de desarrollo Eclipse⁷, integrando sus propias herramientas con aquellas disponibles en la plataforma. AnCora-Pipe se diseñó teniendo en cuenta dos requisitos fundamentales:

1. La posibilidad de ampliación, es decir:
 - a) la posibilidad de configurar y modular el conjunto de atributos y valores, haciendo más fácil al usuario la inclusión o exclusión de los diferentes niveles de análisis lingüístico;
 - b) la implementación de paneles de anotación especializados; y
 - c) la adaptación de herramientas externas para procesos específicos.
2. Gestión multi-alfabética: las herramientas integradas en AnCora-Pipe pueden ser configuradas para trabajar con cualquier tipo de alfabeto.

Las funcionalidades de AnCora-Pipe son principalmente tres:

Funcionalidades

1. la creación de nuevos recursos,
2. su edición y
3. la exportación e importación de los datos a o desde otros entornos de procesamiento.

La creación de nuevos recursos puede realizarse mediante la importación de textos desde formatos externos o mediante la creación de nuevos documentos en la misma plataforma. La edición permite la anotación de corpora y lexicones, así como la modificación de los previamente anotados. La edición se apoya en una serie de interfaces gráficas específicas para cada nivel de análisis lingüístico. Finalmente, AnCora-Pipe proporciona la exportación de los datos para el análisis usando

⁷<http://www.eclipse.org/platform/>

herramientas especializadas como Excel, SPSS, Weka, etc. Un subconjunto de herramientas de importación permite la traslación del formato de AnCora-Pipe (XML, como veremos) a otros formatos genéricos también usados en el análisis y tratamiento de corpus, como el TBF y estructura de dependencias.

Formatos

Los documentos de AnCora-Pipe están en formato XML y usan la codificación UTF-8. Otros formatos se aceptan como entrada pero la salida de la plataforma es siempre en el código UTF-8. Los corpóra y lexicones se almacenan en directorios y carpetas que contienen los documentos, textos en el caso de los corpóra y entradas léxicas en el caso de los lexicones. En AnCora-Pipe cada fichero contiene un solo documento para facilitar y simplificar el manejo de los datos. Se eligió XML como lenguaje de representación porque es un estándar que permite la representación de cualquier tipo de información y admite cualquier tipo de codificación. La codificación, por su parte, es UTF-8 ya que permite la representación de textos en casi todos los sistemas de escritura. Los nodos son las unidades básicas de representación en XML. Se organizan en formato de árbol donde cada nodo puede asociarse a diferentes pares de atributo-valor. En general, la información lingüística se asocia en pares atributo-valor de nodos <word> ('palabra') y <constituent> ('constituyente'). La definición de atributos y sus valores es completamente abierta y adaptable a todo tipo de corpus e información lingüística. Este tipo de organización tan abierta hace más fácil la adaptación de la herramienta para la descripción de una variedad de lenguas y para representar todo tipo de información lingüística.

Perspectivas:*Lexical Annotator for SN*

Las perspectivas son configuraciones gráficas que agrupan un grupo de paneles gráficos para llevar a cabo una tarea concreta. A continuación describimos la perspectiva *Lexical Annotator for SN* que permitió la anotación de los SNs de núcleo deverbal en el corpus AnCora-Es.

En las Figuras 8.5, 8.6, 8.7 y 8.8 vemos el panel de anotación con el que trabajaban los anotadores. En la parte izquierda de las figuras se encuentra la perspectiva *Lexical Annotator for SN*, que según el nodo a anotar, sustantivo (Figuras 8.5 y 8.6) o constituyentes del SN (Figuras 8.7 y 8.8), presenta diferentes botones. La parte central de las figuras alberga al editor, que permite ver el archivo a anotar en tres diferentes vistas: en formato texto (parte superior de las figuras), en forma de estructura sintáctica (parte media de las figuras) y en formato de constituyentes (parte inferior de las figuras). La vista en constituyentes es un panel especialmente diseñado para la anotación de los SNs con la intención de facilitar la visualización de los constituyentes que componen los SNs. Finalmente, en la parte derecha de las figuras se puede hallar o bien la lista de todos los sustantivos por anotar (Figura 8.5) o bien el panel *Lexical Information*, que se activa cuando se selecciona el sustantivo a anotar y que proporciona información sobre el verbo base ('originlexical id'), y ahora también sobre la entrada nominal a la que está asociado el sustantivo (Figura 8.6).

8. ANCORa-ES: VALIDACIÓN MANUAL

La Figura 8.5 muestra el primer estadio del panel de anotación para la validación manual, en cuya parte derecha encontramos todos los sustantivos a anotar, que se obtienen apretando sobre el botón *Search*, ‘búsqueda’ en la perspectiva *Lexical Annotator for SN*. En la vista de constituyentes se facilita la selección del nodo del SN a anotar (el sustantivo o el resto de los constituyentes); el nodo seleccionado se marca con el cambio del sombreado de verde a azul.

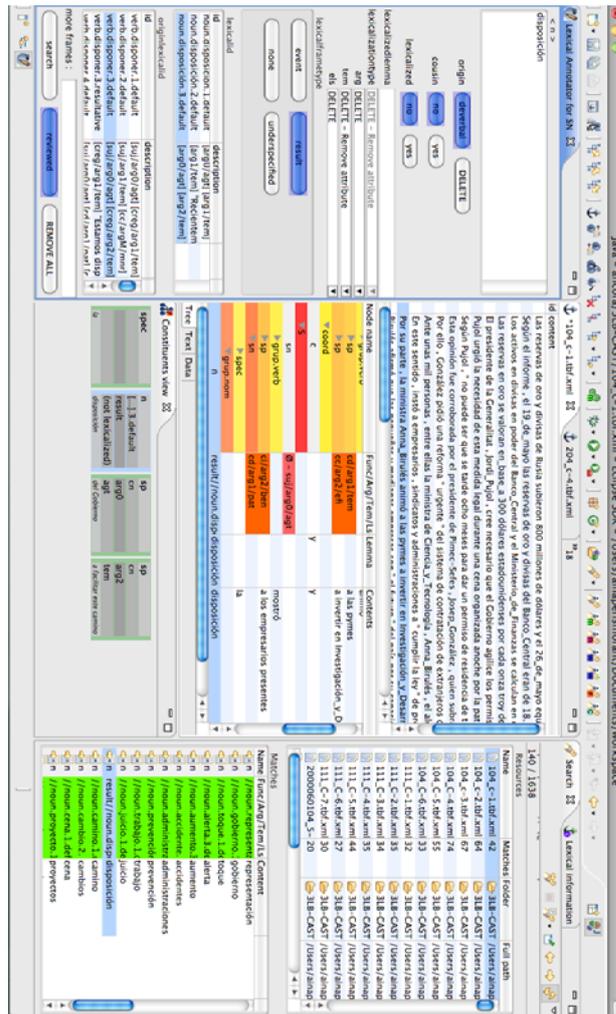


Figura 8.5: AnCora-Pipe para la anotación de los SNs.1

Nos centramos ahora en la parte izquierda de las figuras, en la perspectiva *Lexical Annotator for SN*, que ha sido creada específicamente para la anotación de las nominalizaciones deverbales en el corpus AnCora-Es. Esta perspectiva está diseñada para anotar dos nodos distintos: las nominalizaciones deverbales, núcleo de los SNs, y el resto de constituyentes del SN. En función del nodo seleccionado

en la vista de constituyentes (el sustantivo núcleo o bien alguno de los constituyentes del SN), aparecerá en el panel una información determinada, asociada al nodo escogido. En la Figura 8.6 vemos todos los atributos que se le pueden asociar al sustantivo en esta tarea de anotación. Los atributos que se deben anotar (deverbal, cousin, lexicalized, etc.) están indicados a la izquierda del panel y los valores que cada atributo puede tener se presentan en diferentes botones (los botones relacionados con un mismo atributo son excluyentes, es decir, si se marca uno, los otros quedan excluidos).

En la Figura 8.6, el botón ‘deverbal’ apretado (en azul) indica que la nominalización seleccionada se ha considerado que tiene propiedades deverbales. Si esto no fuera así, el botón ‘delete’ es el que tendría que estar en azul, significando que aquel sustantivo no se ha considerado deverbal (recuérdese el caso de ‘cura’ o ‘establecimiento’). A continuación, el atributo ‘cousin’ debe asociarse a uno de sus dos valores posibles (‘yes’ o ‘no’) y en este caso, el sustantivo ‘disposición’ no se considera *cousin* del verbo ‘disponer’. El siguiente atributo se incorporó al panel de anotación durante el segundo proceso de validación manual, el referido a la denotación. Consiste en señalar mediante los valores ‘yes’ o ‘no’ si el sustantivo forma parte de una construcción lexicalizada. En la Figura 8.6 se observa que el valor marcado para ‘disposición’ en el SN ‘la disposición del Gobierno a facilitar este camino’ es negativo. En el caso de que fuera positivo habría que completar los siguientes dos atributos: <lexicalizedlemma>, donde los anotadores deben especificar la construcción lexicalizada de la que el sustantivo formaría parte, y <lexicalizationtype>, donde los anotadores elegirían entre las seis posibles el tipo de construcción lexicalizada. Le siguen a estos atributos, aquellos que sirven para indicar si el sustantivo tiene un argumento incorporado. Fíjense en que los atributos referentes al argumento (<arg>), papel temático (<tem>) y clase semántica verbal de la que se obtiene dicho argumento (<els>) se anotan en el nodo sustantivo lo que sirve para indicar que el argumento es incorporado. A continuación tenemos el atributo <lexicalframetype>’ al que le corresponden cuatro valores, los tres tipos denotativos y el valor ‘none’ para cuando la nominalización forma parte de una lexicalización que no es de tipo nominal. El valor adecuado se marca y queda señalado en azul. En la Figura 8.6 se muestra que se ha considerado que ‘disposición’ es en este contexto una nominalización resultativa. Los dos últimos atributos relacionan el sustantivo con la entrada nominal y la entrada verbal del verbo base, respectivamente. El atributo ‘lexicalid’ se añadió en última instancia para los casos en los que se debiera corregir, si cabe, la asociación de una ocurrencia del corpus con un determinado sentido nominal. Esta misma información se puede ver en el panel “Lexical Information” de la parte derecha de la figura. En el atributo ‘originlexicalid’ se marca el sentido del verbo del que deriva la nominalización entre todos los sentidos verbales de aquel verbo. En este panel, la información sobre cada sentido verbal no es muy extensa por lo que el anotador

8. ANCORÀ-ES: VALIDACIÓN MANUAL

dispone del panel “Lexical Information” para consultar la entrada verbal correspondiente. El sentido verbal que corresponde a la nominalización se sombrea en azul en el panel “Lexical Annotator for SN”. En la Figura 8.6 se muestra que se ha considerado que ‘disposición’ en este contexto proviene del sentido 3 de ‘disponer’. Finalmente, para indicar que el proceso de validación y anotación se ha finalizado se marca ese nodo como ya revisado, *reviewed*.

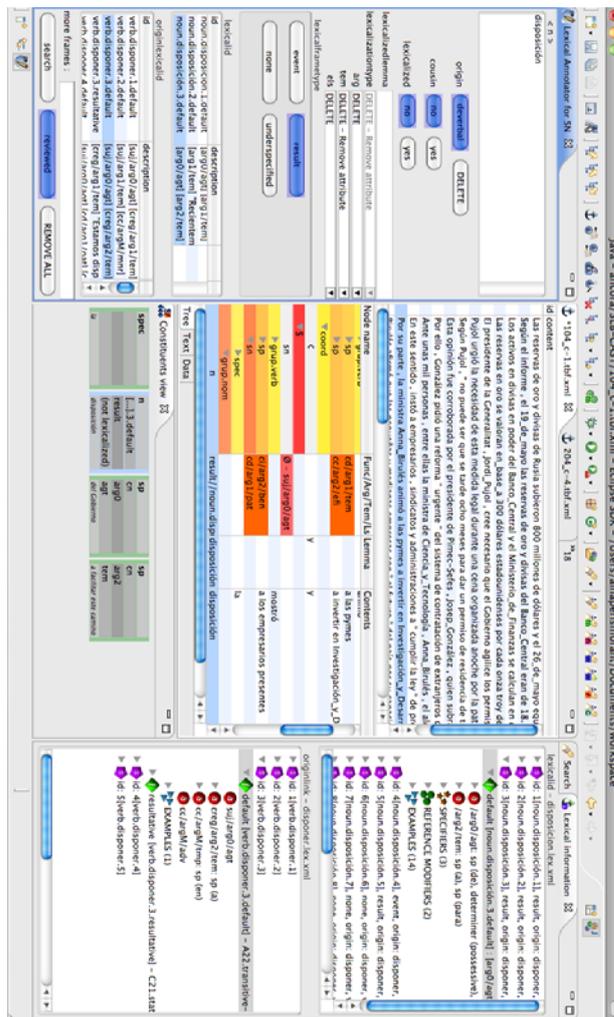


Figura 8.6: AnCora-Pipe para la anotación de los SNs.2

En las Figuras 8.7 y 8.8, vemos los atributos que pueden asociarse al resto de constituyentes del SN: función sintáctica (<func>), argumento (<arg>) y papel temático (<tem>). Recuérdese que para cambiar el nodo a anotar, debe elegirse el constituyente del SN que se quiere anotar en la vista de constituyentes (el constituyente a anotar se sombrea en azul). Los constituyentes que pueden ser ar-

gumentales en un SN pueden ser complementos del nombre (CN), como los SPs, SAs, SAdvS, o no, como los determinates posesivos o los pronombres relativos de genitivo. Los valores del atributo argumento y los del papel temático se corresponden con los vistos en el esquema de anotación presentado en la Sección 3.2. En el SN ‘la disposición del Gobierno a facilitar este camino’, el primer SP es un complemento del nombre (CN) que recibe la interpretación de arg0-agt y así se anota en la perspectiva *Lexical Annotator for SN* (Véase la Figura 8.7). El segundo SP se anota como un CN que recibe la interpretación de arg2-tem (Véase la Figura 8.8).

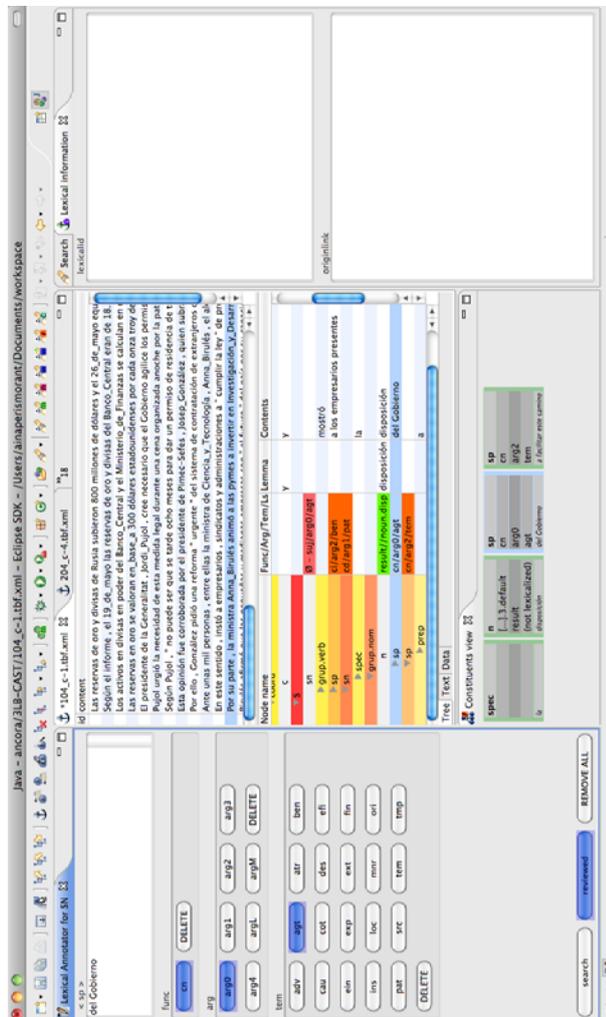


Figura 8.7: AnCora-Pipe para la anotación de los SNs.3

8. ANCORÀ-ES: VALIDACIÓN MANUAL

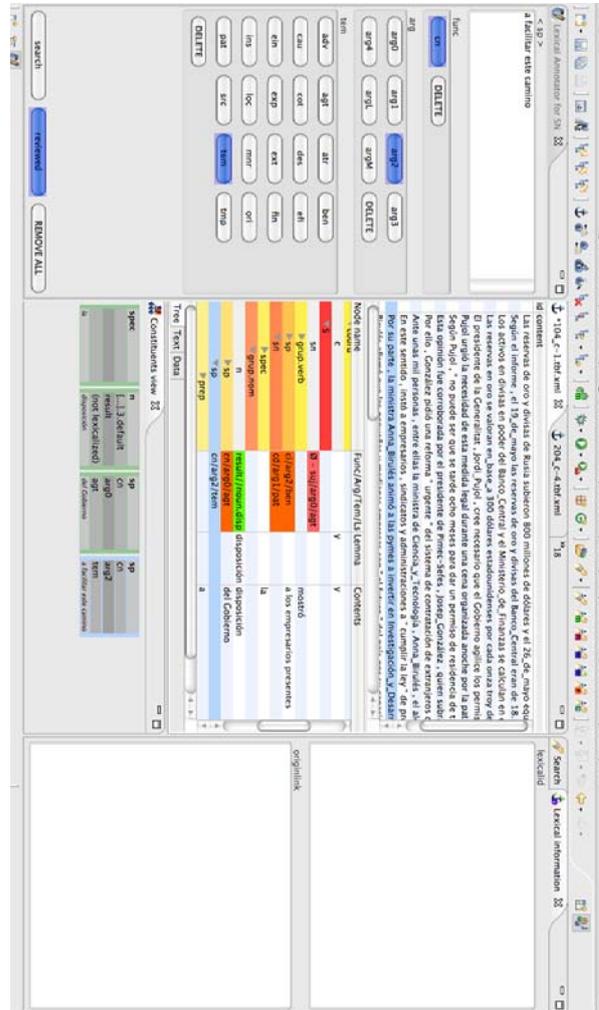


Figura 8.8: AnCora-Pipe para la anotación de los SNs.4

El uso de AnCora-Pipe como herramienta de anotación ha facilitado la tarea en dos sentidos: en primer lugar, ha minimizado errores puesto que los anotadores no tenían que escribir nada excepto en el caso del atributo <lexicalizedlemma> que no es habitual. En segundo lugar, ha ahorrado tiempo en la tarea de anotación porque la anotación de cada atributo consistía en un clic, en el caso de que los valores fueran reducidos, o como máximo dos, en el caso de que hubiera que desplegar las opciones (un clic) y elegir una (segundo clic).

8.4. Conclusiones: AnCora-Es-v3

Los dos procesos de validación manual descritos han dado lugar al corpus AnCora-Es anotado con la estructura argumental y la denotación de las nominalizaciones deverbales. En la Figura 8.9 podemos observar cómo queda anotado en el corpus el SN ‘la ampliación de ocho a doce meses el periodo de referencia’.

```

<sn>
  <spec gen="f" num="s">
    <d gen="f" lem="el" num="s" pos="da0fs0" postype="article" wd="la"/>
    <grup.nom gen="f" num="s">
      <n gen="f" lem="ampliación" denotationtype="event" num="s" originlexicalid="verb.ampliar.1.default" pos="ncfs000"
      postype="common" sense="16:00235235" wd="ampliación"/>
      <sp arg="arg3" func="cn" tem="ein" >
        <prep>
          <s lem="de" pos="sps00" postype="preposition" wd="de"/>
        </prep>
        <sn>
          <grup.nom>
            <p gen="c" lem="ocho" num="p" pos="pn0cp000" postype="numeral" wd="ocho"/>
            <sp arg="arg4" func="cn" tem="efi" >
              <prep>
                <s lem="a" pos="sps00" postype="preposition" wd="a"/>
              </prep>
              <sn>
                <spec gen="m" num="p">
                  <d gen="c" lem="doce" num="p" pos="dn0cp0" postype="numeral" wd="doce"/>
                  <grup.nom gen="m" num="p">
                    <n gen="m" lem="mes" num="p" pos="ncmp000" postype="common" sense="16:10919146" wd="meses"/>
                    <sp arg="arg1" func="cn" tem="tem" >
                      <prep>
                        <s contracted="yes" gen="m" lem="del" num="s" pos="spcms" postype="preposition" wd="del"/>
                      </prep>
                      <sn>
                        <grup.nom gen="m" num="s">
                          <n gen="m" lem="periodo" num="s" pos="ncms000" postype="common" sense="16:10843624" wd="periodo"/>
                          <sp>
                            <prep>
                              <s lem="de" pos="sps00" postype="preposition" wd="de"/>
                            </prep>
                            <sn>
                              <grup.nom gen="f" num="s">
                                <n gen="f" lem="referencia" denotationtype="none" num="s" originlexicalid="verb.referir.2.default" pos="ncfs000"
                                postype="common" sense="16:05417191" wd="referencia"/>
                              </grup.nom>
                            </sn>
                          </sp>
                        </grup.nom>
                      </sn>
                    </sp>
                  </grup.nom>
                </sn>
              </sp>
            </grup.nom>
          </sn>
        </sp>
      </n>
    </grup.nom>
  </spec>

```

Figura 8.9: Ejemplo de anotación de ‘ampliación’ en AnCora-Es

En este ejemplo podemos observar que el sustantivo ‘ampliación’ tiene en este SN tres argumentos: un argumento tema (arg1-tem), un argumento estado inicial (arg3-ein) y un argumento estado final (arg4-efi). Una estructura oracional equivalente sería ‘se amplía el periodo de referencia de ocho a doce meses’. La denotación es eventiva y así queda reflejado en el atributo <denotationtype> (<denotationtype=event>). Como los criterios de pluralidad, determinación y complementación no arrojaban luz suficiente, entonces fue necesario consultar la clase semántica del sentido nominal atribuido a la nominalización en el atributo <originlexicalid>. Dado que el sentido verbal pertenece a la clase de los logros había que observar qué argumentos se realizaban sintácticamente. En este caso el arg1 está explícito y aunque el arg2 no, consideramos que el arg3 y arg4 explícitos apoyan la lectura eventiva de la misma forma que lo haría un arg2.

El enriquecimiento de AnCora-Es con la anotación de la estructura argumental y denotación de las nominalizaciones deverbales, es una de las contribuciones importantes de nuestro trabajo. Un corpus de dichas características puede ser útil en investigaciones lingüísticas sobre las nominalizaciones deverbales del español ya que proporciona una gran variedad de casos y datos reales, pero además también puede ser un recurso que sirve para desarrollar sistemas automáticos de SRL nominal para el español u otras herramientas, el clasificador ADN es un ejemplo de utilización del corpus como recurso de aprendizaje. Además, la anotación del corpus nos ha permitido inducir el léxico AnCora-Nom. Describimos el proceso de inducción y el propio léxico en el capítulo siguiente.

CAPÍTULO 9

ANCORA-NOM: UN LÉXICO DE NOMINALIZACIONES DEVERBALES

En este capítulo presentamos el léxico AnCora-Nom, un léxico de nominalizaciones deverbales del español. Actualmente contiene 1.655 entradas que se corresponden con los diferentes lemas de nominalizaciones deverbales que aparecen en el corpus AnCora-Es. En AnCora-Nom cada sentido nominal se asocia con un tipo denotativo (evento, resultado o subespecificado) y además para cada sentido se anotan sus argumentos y correspondientes papeles temáticos. Una particularidad de este léxico, en contraste con la mayoría de los reseñados en el Capítulo 2 de esta tesis, es que este se genera de manera automática a partir de la información anotada en el corpus AnCora-Es (Capítulo 8) y los otros se construyen de manera manual. A continuación nos centramos en el proceso de elaboración del léxico AnCora-Nom (Sección 9.1). En la siguiente sección se detalla la información lingüística codificada en AnCora-Nom (Sección 9.2). A continuación se ofrecen algunos datos cuantitativos de interés sobre este léxico (Sección 9.3). Finalmente, el capítulo termina con unas conclusiones (Sección 9.4)

9.1. Proceso de creación del léxico AnCora-Nom

La creación de un léxico que recoge las propiedades de las nominalizaciones deverbales (denotación y estructura argumental) nos pareció necesario desde el inicio de nuestro trabajo. Con un léxico de estas características es posible el análisis de las propiedades combinatorias de las nominalizaciones, según el tipo denotativo y también en función de su estructura argumental; es decir, disponer de este tipo de léxico constituye una herramienta muy útil para el análisis lingüístico

porque permite identificar los rasgos distribucionales más prototípicos e importantes para este tipo de predicado. Además, puede ser también un recurso útil para el PLN, como en nuestro caso, en el que el clasificador ADN utiliza AnCora-Nom para extraer algunos de los atributos que emplea.

A continuación, describimos el proceso de elaboración del léxico AnCora-Nom, construido de manera incremental. Esto es, existen tres versiones distintas de este léxico en las que a medida que se avanza se incrementa o bien el número de nominalizaciones o bien el tipo de información declarado. Cada una de las tres versiones corresponde a una fase distinta del proceso de investigación.

AnCora-Nom-v1

Desde un primer momento se juzgó necesario la creación de un léxico que recogiera las propiedades de las nominalizaciones deverbales (denotación y estructura argumental) que nos permitiera obtener datos estadísticos de este tipo de predicados. Por eso, durante el estudio empírico basado en corpus que se llevó a cabo al principio de este trabajo, se creó manualmente la primera versión de AnCora-Nom. AnCora-Nom-v1 constaba de un total de 817 entradas nominales, correspondientes a los 817 lemas de nominalizaciones deverbales que se encontraban en el subconjunto de 100.000 palabras de AnCora-Es sobre el que se realizó el mencionado estudio. Esta primera versión contenía información sobre el tipo denotativo de las nominalizaciones, a partir del cual se establecían los distintos sentidos (los diferentes tipos denotativos, constituían sentidos distintos), y para cada sentido se asociaba la estructura argumental, es decir, se recogía qué argumentos se realizaban para esa nominalización y mediante qué constituyentes, y también características morfosintácticas como el tipo de determinante o la pluralización, además de los ejemplos del corpus que mostraban todas estas características. A partir de este léxico desarrollado manualmente se creó la primera versión del clasificador ADN (-v1).

AnCora-Nom-v2

Los experimentos con ADN-v1 revelaron que la información de la estructura argumental era muy importante para la clasificación de la denotación, por lo que si queríamos anotar la denotación en el conjunto de AnCora-Es debíamos tener información sobre la estructura argumental. Tras la anotación y posterior validación manual de los argumentos de todas las nominalizaciones del corpus AnCora-Es (un total de 1.655 lemas y sus 23.431 ocurrencias correspondientes), se generó, esta vez automáticamente, una nueva versión del léxico AnCora-Nom. AnCora-Nom-v2 constaba, por lo tanto, de un total de 1.655 entradas léxicas, correspondientes a los lemas de AnCora-Es. Esta versión tiene dos particularidades: en primer lugar, las entradas ya existentes, las 817 de AnCora-Nom-v1 se modificaron en dos sentidos: 1) se aumentó el número de sentidos de cada entrada ya que además del tipo denotativo previamente asociado se determinaron sentidos nominales en función del sentido verbal asociado a la nominalización (recuérdese que esto se validaba en el primer proceso de validación manual, junto a la estructura argumental) y 2) se modificaron algunos atributos (estructura argumental, tipos de

determinantes, plural) a partir de la información obtenida automáticamente de las ocurrencias del corpus correspondiente a los 817 lemas en el resto del corpus; estas ocurrencias también se añadieron como nuevos ejemplos de estas entradas. En segundo lugar, las 838 entradas nominales restantes, correspondientes a los lemas que no se encontraban en el subconjunto inicial de 100.000 palabras de AnCora-Es, se generaron automáticamente a partir de los datos previamente anotados en el corpus y, por lo tanto, no contenían información sobre el tipo denotativo, que aún debía ser anotado. Es decir, en estas 838 entradas los sentidos estaban únicamente determinados por el distinto sentido del verbo base y la información sobre la estructura argumental y las propiedades morfosintácticas se obtenían automáticamente del corpus. A partir de AnCora-Nom-v2 y del corpus AnCora-Es, el clasificador ADN-v2 (la versión de ADN que adapta el modelo de sentidos –aprendido a partir de las 100.000 palabras del corpus AnCora-Es y de AnCora-Nom-v1– a un modelo de lemas, se anota automáticamente la denotación de todas las ocurrencias del corpus AnCora-Es. Tras la validación manual de esta información, se genera la última y definitiva versión del léxico, AnCora-Nom-v3, cuyo proceso de inducción se detalla en la siguiente Subsección 9.1.1. Por lo tanto, el proceso de elaboración de AnCora-Nom¹ ha sido un proceso incremental, tal y como queda reflejado en la Figura 9.1.

AnCora-Nom-v3

La versión final de AnCora-Nom (AnCora-Nom-v3) ha sido creada automáticamente a partir del corpus AnCora-Es. Este hecho lo distingue claramente de otros léxicos similares descritos en la Sección 2.2.1, como el de NomLex (Macleod et al., 1998) para el inglés, FrameNet para el inglés (Ruppenhofer et al., 2006), FrameNet para el alemán (Burchardt et al., 2009), FrameNet para el español (Subirats, 2009), FrameNet para el japonés (Ohara, 2009), *Essex Data-Base of Russian Nominalizations* para el ruso (Spencer and Zaretskaya, 1999) y Noma-ge para el francés (Balvet et al., 2010). Otro hecho que distingue a AnCora-Nom-v3 es que es el único léxico para el español de nominalizaciones deverbales que contiene información sobre la denotación y la estructura argumental. En el proyecto FrameNet del español (Subirats, 2009) encontramos 1.200 entradas léxicas entre sustantivos, adjetivos y verbos, por lo que si se quieren estudiar las nominalizaciones deverbales del español, AnCora-Nom-v3 parece un buen recurso del que partir. A continuación vemos cómo se ha obtenido.

9.1.1. Proceso de extracción

La metodología empleada para construir el léxico AnCora-Nom consiste en aprovechar la información anotada en el corpus AnCora-Es. Como se ha mencionado anteriormente, el corpus AnCora-Es contiene 500.000 palabras anotadas

¹A partir de ahora, nos referiremos a AnCora-Nom para hablar de la última y definitiva versión.

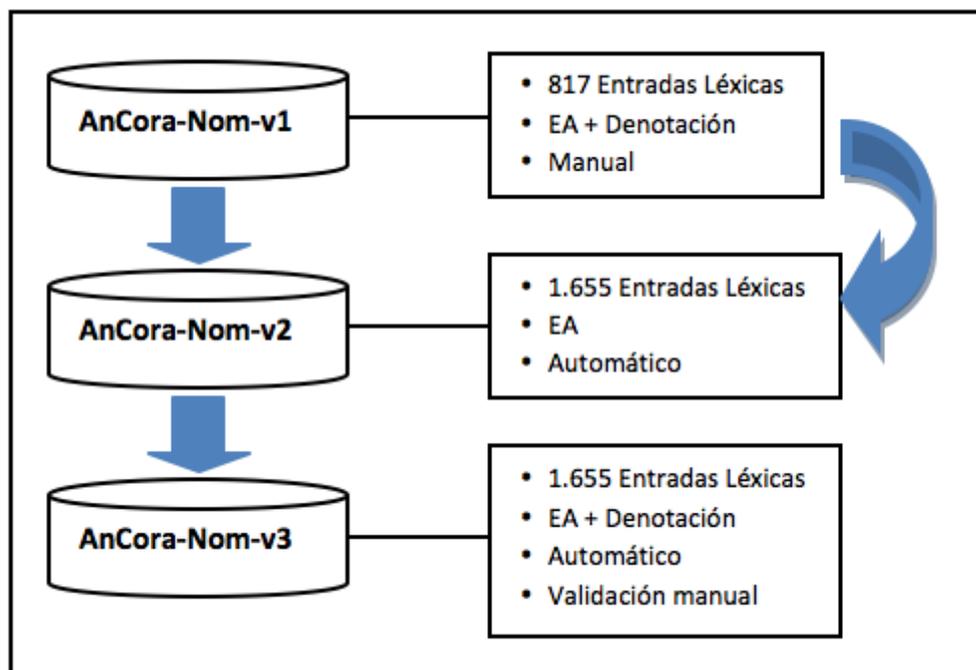


Figura 9.1: Proceso de elaboración incremental del léxico AnCora-Nom

a diferentes niveles lingüísticos a los que cabe añadir la anotación de las nominalizaciones deverbales (23.431 ocurrencias) que hemos descrito en los capítulos anteriores. Concretamente, se ha anotado la estructura argumental de dichas nominalizaciones, su interpretación semántica (evento, resultado, subespecificado) y si la nominalización forma parte de una construcción lexicalizada y de qué tipo. El léxico AnCora-Nom se ha derivado no solo de la información semántica estrictamente relacionada con las nominalizaciones sino también de la información morfológica y sintáctico-semántica previamente anotada en el corpus. Describimos con detalle en esta subsección el proceso de extracción de las entradas léxicas a partir del corpus.

Como ya se ha avanzado, la obtención de las entradas léxicas correspondientes a las nominalizaciones deverbales se ha realizado de manera automática, la versión final de AnCora-Nom se ha obtenido a partir de la información anotada en las 23.431 ocurrencias del corpus AnCora-Es, que corresponden a un total de 1.655 lemas. Para cada uno de estos lemas, se ha creado una entrada léxica que responde a una estructura jerárquica fija, constituida por diferentes nodos, a los que se les pueden asociar uno o más atributos, con su correspondiente valor. Las entradas léxicas, como los documentos del corpus AnCora-Es, se representan en formato XML y se codifican en UTF-8. A continuación se muestra la estructura básica de

las entradas léxicas de AnCora-Nom (Figura 9.2), donde los principales nodos se marcan en **negrita** y los atributos se señalan mediante el subrayado.

```
<?xml version="3.0" encoding="UTF-8"?>
<lexentry lemma=" " lng=" " origin=" " type=" ">
  <sense alternativelemma=" " cousin=" " denotation=" " id=" " lexicalizationtype=""
  lexicalized=" " originlemma=" " originlink=" " wordnetsynset=" ">
    <frame appearsinplural=" " type=" ">
      <argument argument=" " thematicrole=" ">
        <constituent1 frequency=" " preposition=" " type=" "/>
        <constituentn frequency=" " posttype=" " type=" "/>
      </argument>
      <reference_modifier >
        <constituent1 frequency=" " preposition=" " type=" "/>
        <constituentn frequency=" " preposition=" " type=" "/>
      </reference_modifier>
      <specifiers>
        <constituent1 frequency=" " posttype=" " type=" "/>
        <constituentn frequency=" " posttype=" " type=" "/>
      </specifiers>
      <examples>
        <example file=" " nodepath=" " sentencenodepath=" ">
          </example>
        </examples>
      </frame>
    </sense>
  </lexentry>
```

Figura 9.2: Estructura de entrada léxica de AnCora-Nom

En la Figura 9.2 se observa la estructura jerárquica básica de una entrada léxica de AnCora-Nom. A continuación detallamos cada uno de los nodos que la componen y su organización jerárquica e introducimos los atributos asociados a cada uno de ellos. En la Sección 9.2 se describen los valores posibles de cada uno de los atributos con detalle.

El nodo **<lexentry>** es el nodo raíz para cada una de las entradas léxicas. En este nodo se asocian los atributos referidos al lema que se representa en aquella entrada léxica (“lemma”), a la lengua representada en la entrada (“lng”), al origen del lema, a si aquel lema se genera a partir de un proceso derivativo (“origin”) y a la categoría sintáctica de dicho lema (“type”). Cada entrada léxica puede tener uno o más sentidos.

El nodo **<sense>** agrupa a todos los marcos o esquemas sintáctico-semánticos nominales de un mismo sentido nominal (uno o más marcos). En este nodo

Nodos y atributos

se asocian los atributos que indican si la nominalización es una nominalización *cousin* (“cousin”), el tipo denotativo de la nominalización (“denotation”), el identificador del sentido (“id”), si aquel sentido de la nominalización está lexicalizado (“lexicalized”), el lema verbal con el que se relaciona la nominalización (“originlemma”), el sentido concreto del lema verbal relacionado con la nominalización (“originlink”) y los synsets de WordNet asociados a dicho sentido nominal (“wordnetsynset”). En el caso de que el sentido nominal sea lexicalizado se activan dos atributos opcionales, que son: la construcción lexicalizada de la que forma parte la nominalización (“alternativelemma”) y el tipo de lexicalización (“lexicalizationtype”).

El nodo <**frame**> hace referencia al marco o marcos nominales que un sentido puede tener, es decir, a los distintos esquemas sintáctico-semánticos. Por marco nominal entendemos el nivel nominal que se corresponde con el nivel de marco verbal en las entradas verbales que se establecen según las alternancias de diátesis en las que participa un verbo. Es importante preservar este nivel de marco porque es en este nivel donde se especifica la estructura argumental de los verbos y por tanto, también es en el correspondiente nivel nominal donde se especifica la estructura argumental de las nominalizaciones. Los atributos asociados a este nodo son la aparición en plural o no de la nominalización (“appearsinplural”) y el marco verbal concreto del que deriva la nominalización (“type”). Este nodo además está formado por otros nodos, todos en el mismo nivel de jerarquía que especifican la estructura argumental de las nominalizaciones <**argument**>, indican si la nominalización tiene complementos no argumentales <**reference-modifier**>, el tipo de determinante que especifica la nominalización <**specifiers**> y los ejemplos asociados a dicho marco <**examples**>. Los primeros dos nodos (<argument> y <reference-modifier>) son opcionales, puede darse el caso de marcos nominales correspondientes a nominalizaciones que no tienen argumentos o que no tienen complementos no argumentales. Cada uno de estos nodos, a su vez, tiene atributos asociados.

En el nodo <**argument**> se especifican los diferentes argumentos asociados al marco nominal. Para cada argumento se detalla mediante atributos la posición argumental (“argument”) y el papel temático (“thematicrole”). También para este nodo se especifican los diferentes constituyentes (“type”) que realizan dichos argumentos y la frecuencia en que lo realizan, es decir, el número de veces (“frequency”). En el caso de que el constituyente argumental sea un SP se incluye un atributo de constituyente opcional, el tipo de preposición del SP (“preposition”), y si es un determinante posesivo se indica además del tipo, el subtipo de determinante (“postype”).

En el nodo <**reference-modifier**> se especifican los diferentes constituyentes que representan los complementos de las nominalizaciones no argumentales.

Como en el caso de los argumentos, se especifica el tipo de constituyente (“type”) y su frecuencia (“frequency”). También si el constituyente es un SP se incluye al atributo opcional (“preposition”)².

En el nodo <**specifiers**> se representa cómo se especifica la nominalización, es decir, mediante qué constituyentes se pueden especificar las nominalizaciones. Para cada uno de ellos se marca el tipo de constituyente (“type”), el subtipo en el caso de que el constituyente sea un determinante (“postype”) y la frecuencia con la que dicho constituyente especifica la nominalización (“frequency”).

Finalmente, en el nodo <**examples**> se concretan cada uno de los ejemplos del corpus AnCora-Es asociados a ese marco nominal. Para cada ejemplo <example> se declaran los atributos referentes al fichero del corpus en el que se encuentra el ejemplo (“file”), al camino en el fichero hasta llegar a la nominalización correspondiente (“nodepath”) y al camino en el fichero hasta llegar a la oración que contiene la nominalización, es decir, el número de oración empezando a partir de 0 (“sentencenodepath”).

Tomando como punto de partida esta estructura de entrada léxica, la generación de las diferentes entradas consiste básicamente en dos procesos automáticos a partir de la consulta de todas las ocurrencias de cada lema:

Generación de las entradas léxicas

- Determinar el número de sentidos diferentes que puede tener una nominalización.
- Extraer para cada sentido los atributos asociados a este nodo, así como establecer el número de marcos nominales que contiene dicho sentido y sus correspondientes atributos. En otras palabras, se extraen los valores de todos los atributos posibles del nodo sentido, es decir, los atributos concretos para la caracterización de dicho sentido.

Para establecer los sentidos de las nominalizaciones a partir de las ocurrencias se tiene en cuenta: el tipo denotativo, el sentido del verbo base y si la nominalización forma parte de una construcción lexicalizada. En concreto, los criterios seguidos son:

Delimitación de sentidos

- Si la ocurrencia de la nominalización forma parte de una construcción lexicalizada, entonces constituye un sentido por sí misma. En aquel sentido solo se incluyen las ocurrencias de nominalizaciones que respondan a la misma construcción lexicalizada, es decir, aquellas ocurrencias de nominalizaciones que compartan un mismo “alternativelemma”. Por ejemplo, la

²Dado que no existen determinantes que sean complementos del nombre, en este nodo no encontramos como atributo opcional posible el subtipo de determinante (“postype”).

nominalización ‘caída’ cuando se encuentra en la construcción ‘en caída libre’ (1) constituye un sentido diferente de la misma nominalización en una oración como (2). Y además, solo ocurrencias de ‘caída’ en dicha construcción podrán adherirse a este sentido lexicalizado.

- (1) El laborartorio se encontrará [en **caída** libre]SP.
- (2) El aumento del paro provocó [una fuerte **caída** del índice Nikkei de la Bolsa de Tokio]SN.

- Las ocurrencias de nominalizaciones pertenecientes al mismo lema que cumplan dos condiciones, tener el mismo tipo denotativo y ser derivadas del mismo sentido verbal, se agrupan bajo un mismo sentido nominal. La primera condición se obtiene del atributo del corpus “denotationtype”³ y la segunda condición se obtiene a partir del atributo del corpus “originlexicalid” consultando el valor (*verb.adelantar.3.default*) hasta el número, lo que viene a continuación es el marco verbal, que se utilizará para establecer los marcos nominales. Por ejemplo, el sustantivo ‘adelanto’ en (3) se ha marcado como un evento en el corpus (fíjense que la nominalización podría ser sustituida por una cláusula verbal, ‘para conseguir que se adelanten las elecciones’), mientras que la misma nominalización en (4) ha sido marcada como un resultado (el plural, el adjetivo relacional y el hecho de que no pueda ser equivalente a una cláusula verbal apoyan esta clasificación semántica), y a pesar de que ambas están asociadas al mismo sentido verbal (<originlexicalid=“verb.adelantar.3.default”>), constituyen sentidos diferentes de la entrada léxica porque tienen distinto tipo denotativo. También pertenecen a dos sentidos nominales diferentes de la entrada léxica de ‘organización’ las ocurrencias de (5) y (6), pero en este caso porque se derivan de dos sentidos diferentes de ‘organizar’ a pesar de que la denotación sea la misma. El sustantivo ‘organización’ en (5) se deriva del primer sentido de ‘organizar’ (<originlexicalid=“verb.organizar.1.default”>), que significa según el *Diccionario de la Real Academia Española* (Real Academia de la Lengua Española, 2012) “establecer o reformar algo para lograr un fin, coordinando las personas y los medios adecuados. En cambio ‘organización’ en (6) se deriva del segundo sentido de ‘organizar’ (<originlexicalid=“verb.organizar.2.default”>), que significa según el *Diccionario de la Real Academia Española* “Poner algo en orden”.

³En el corpus, los atributos asociados a las nominalizaciones deverbales, así como sus posibles valores, aparecen en inglés. Por esta razón, los ejemplos que aparecen a partir de ahora tendrán en inglés el nombre de los atributos y valores, para respetar la nomenclatura del corpus.

9. ANCORAS-NOM: UN LÉXICO DE NOMINALIZACIONES DEVERBALES

- (3) El Shas respalda a los partidos de la oposición israelí para conseguir [el **adelanto** <denotationtype="event"> <originlexicalid="verb.adelantar.3.default"> de las elecciones]SN.
- (4) Al borde del siglo XX crean Daimler, un automóvil que reunía todos [los **adelantos**<denotationtype="result"> <originlexicalid="verb.adelantar.3.default"> técnicos del momento]SN.
- (5) Tres detalles certifican que el motociclismo español no estaba yermo: la presencia de un nutrido grupo de pilotos muy prometedores en la categoría reina; la potente aportación económica de patrocinadores y [la **organización** <denotationtype="event"> <originlexicalid="verb.organizar.1.default"> este año de tres grandes premios]SN.
- (6) [La **organización**<denotationtype="event"> <originlexicalid="verb.organizar.2.default"> urbanística de la ciudad]SN en el siglo XIX la llevó a cabo Ildefons Cerdà.

Los sentidos nominales establecidos a partir de las 23.431 ocurrencias anotadas en el corpus ascienden a un total de 3.094 para las 1.655 entradas léxicas, por lo que el promedio de sentidos por lema es de 1,87. El número de entradas monosémicas son 883, lo que supone el 53 % de las entradas. La mayoría de estas entradas monosémicas se corresponden a sentidos resultativos (764 de las 883, es decir, el 86 % de las entradas monosémicas), seguidas muy de lejos por los sentidos monosémicos eventivos (78 de las 883, es decir, un 9 %) y subespecificados (37 de las 883, es decir, un 4 %) y los sentidos monosémicos correspondientes a las lexías no nominales (4 de las 883, es decir, un 1 %). Entre las entradas polisémicas (772, el 47 % restante), la mayoría tiene dos sentidos (407), lo que supone el 52 % de las entradas polisémicas; entradas polisémicas con tres sentidos hay 224 (29 %), con cuatro 78 (10 %), con cinco 31 (4 %) y con más de cinco 35 (5 %). La entrada léxica con más sentidos es la de 'cuenta' que tiene 13 sentidos, en gran parte porque tiene un importante número de sentidos lexicalizados.

Una vez que se establecen los sentidos nominales, para cada uno de ellos se extraen del corpus los atributos asociados al nodo sentido <sense> (que se detallan en la Sección 9.2.2). Además, para cada sentido se deben establecer uno o más marcos (*frame*), dependiendo del marco verbal concreto del que deriva la nominalización. El marco del verbo base se extrae del atributo del corpus <originlexicalid>, cuya última parte del valor (después del número de sentido) especifica el marco verbal (*verb.adelantar.3.default*). En total, existen 3.204 marcos nominales diferentes, un promedio de 1,1 marcos por sentido, es decir, que en la gran mayoría de los casos para cada sentido solo se contempla un marco. Esto es así porque las nominalizaciones tienden a derivar del marco verbal no marca-

Extracción de atributos

do (“default”) y no de marcos verbales marcados como los pasivos (“passive”) o inacusativos (“unaccusative”).

Como hemos visto en la Figura 9.2 en el nivel del marco nominal se representan atributos tan importantes como la estructura argumental, los complementos no argumentales, el tipo de determinante, el atributo que indica si ese sentido nominal puede o no aparecer en plural, el tipo de marco nominal y los ejemplos del corpus asociados al marco nominal en cuestión. En los tres primeros atributos se marca la frecuencia, esto es, el número de veces que un determinado constituyente es argumento o complemento no argumental de dicha nominalización o que un determinante aparece como especificador de la nominalización. Para establecer la frecuencia es necesario consultar todas las ocurrencias del corpus de dicha nominalización y poder contar el número de veces de cada fenómeno. El atributo de la pluralidad no se codifica en el léxico hasta que se han consultado todas las ocurrencias de un mismo marco: si una aparece en plural, el valor es positivo.

Este proceso de extracción no ha dado prácticamente lugar a errores en la generación del léxico AnCora-Nom. De hecho, revisando manualmente el léxico se han encontrado algunas entradas léxicas con errores (53 entradas) que se caracterizaban por tener más sentidos de los necesarios. Sin embargo, estos errores no se dan por un fallo en el proceso automático de extracción sino que provienen de errores en la anotación del corpus. En la mayoría de los casos, se trataba de sentidos de más porque la nominalización había sido asociada a más de un sentido verbal erróneamente, seguramente debido a una mala elección del sentido verbal concreto en el desplegable de AnCora-Pipe (véase la Sección 8.3). En otros casos, aunque menos, se debía a un error en la asociación de la denotación a una ocurrencia que había generado un sentido de más.

En la siguiente sección, detallamos la información que contiene cada uno de los atributos que definen la entrada léxica. Una vez creadas las 1.655 entradas, se eliminó la información del corpus relacionada con la denotación y las estructuras lexicalizadas ya que así se evitaba duplicar la misma información en dos recursos, y en su lugar, se dejó un puntero en cada ocurrencia a su correspondiente entrada nominal (sentido y marco) en la que se declara dicha información. Además, cada entrada léxica nominal está a su vez relacionada con la correspondiente verbal, por lo que AnCora-Nom y AnCora-Verb son recursos completamente relacionados.

9.2. AnCora-Nom

En esta sección describimos con detalle la información especificada en los atributos de las entradas nominales de AnCora-Nom (Peris and Taulé, 2011a). La Figura 9.3 y la Figura 9.4 nos sirven para ejemplificar cada uno de los atributos que describimos. Por razones expositivas organizamos los atributos en tres grupos en

función del nodo al que se asocian: atributos a nivel de entrada léxica <lexentry> (Subsección 9.2.1), atributos a nivel de sentido <sense> (Subsección 9.2.2) y atributos a nivel de marco nominal <frame> (Subsección 9.2.3).

9.2.1. Atributos a nivel de entrada léxica

Los atributos a nivel de entrada léxica no se obtienen a partir del corpus, sino que se generan automáticamente con sus respectivos valores. Sirven para documentar el tipo de entrada léxica y son los siguientes:

- **Lema** [“**lemma**={**lema₁**, **lema_n**}”]⁴. En este atributo se indica como valor el lema correspondiente a la entrada léxica. Se ha considerado como lema la forma singular del sustantivo. En la Figura 9.3 el valor para este atributo es la nominalización ‘aceptación’ (lemma=“aceptación”), mientras que en la Figura 9.4 el valor es la nominalización ‘golpe’ (lemma=“golpe”).
- **Lengua** [“**lng**={**es**, **ca**}”]. Este atributo codifica la lengua representada en la entrada léxica. Los recursos AnCora trabajan tanto en español como en catalán, por lo que los valores de este atributo son “es” para el español (lng=“es”) y “ca” para el catalán (lng=“ca”). Actualmente, AnCora-Nom solo trata nominalizaciones deverbales del español, por lo que en las Figuras 9.3 y 9.4 el valor para este atributo es siempre “es” (lng=“es”). En un futuro próximo las nominalizaciones del catalán también se tendrán en cuenta.
- **Origen** [“**origin**={**deverbal**, **deadjectival**}”]. Este atributo indica el tipo de palabra de la que deriva la nominalización. Actualmente, AnCora-Nom solo contiene nombres deverbales pero en un futuro incluirá otro tipo de nominalizaciones como las deadjetivales (‘sutillieza’). En las Figuras 9.3 y 9.4, el valor para este atributo es “deverbal” (origin=“deverbal”), lo que significa que estas entradas tratan nominalizaciones derivadas de verbos.
- **Tipo** [“**type**={**noun**, **verb**}”]. Este atributo identifica el tipo de palabra (entiéndase, categoría sintáctica) representado en aquella entrada léxica. Los recursos AnCora también trabajan con verbos, por lo que los valores para este atributo son “verb” en el caso de entradas léxicas verbales y “noun” en las entradas léxicas nominales. Por lo tanto, en AnCora-Nom todas las entradas tendrán el valor “noun” para este atributo, como ejemplifican las Figuras 9.3 y 9.4.

⁴Para cada atributo describimos entre corchetes el nombre del lema, tal y como se representa en la entrada léxica, y sus posibles valores, que aparecen entre llaves.

```

<?xml version="1.0" encoding="UTF-8"?>
<lexentry lemma="aceptación" lng="es" origin="deverbal" type="noun">
  <sense cousin="no" denotation="result" id="1" lexicalized="no"
  originlemma="aceptar" originlink="verb.aceptar.1"
  wordnetsynset="16:00117820+16:10039397">
    <frame appearsInplural="no" type="default">
      <argument argument="arg0" thematicrole="agt">
        <constituent frequency="1" preposition="de" type="sp"/>
        <constituent frequency="1" type="s.a"/>
      </argument>
      <specifiers>
        <constituent frequency="1" posttype="article" type="determiner"/>
        <constituent frequency="1" type="void"/>
      </specifiers>
      <examples>
        <example file="CESS-CAST-P/141_19981202.tbf.xml"
        nodepath="4.5.3.2.1.0" sentencenodepath="4">Para el realizador y guionista , el
        protagonista masculino , Stéphane , " es muy interesante porque Ø encarna la
        tolerancia , aceptación de los demás . </example>
        ... </example></frame></sense>
      <sense cousin="no" denotation="event" id="2" lexicalized="no"
      originlemma="aceptar" originlink="verb.aceptar.1"
      wordnetsynset="16:00117820">
        <frame appearsInplural="no" type="default">
          <argument argument="arg1" thematicrole="pat">
            <constituent frequency="2" preposition="de" type="sp"/>
            <constituent frequency="1" posttype="possessive" type="determiner"/>
          </argument>
          <specifiers>
            <constituent frequency="2" posttype="article" type="determiner"/>
          </specifiers>
          <examples>
            <example file="CESS-CAST-A/11714_20000314.tbf.xml"
            nodepath="7.4.1.1.1.3.2.1.2.0" sentencenodepath="7">En el marco de esta
            estrategia marcada por la prudencia , el PP esperará a los movimientos que haga el
            consejero de Economía , desde la determinación de que cualquier apoyo popular
            dependerá de la " capacidad de diálogo y de llegar a acuerdos " que muestre CiU y
            de la aceptación de nuestra capacidad de influencia </example>
            </examples> </frame> </sense>
          </lexentry>

```

Figura 9.3: Entrada léxica de ‘aceptación’

9.2.2. Atributos a nivel de sentido

En esta subsección se detallan los atributos asociados al sentido. Recuérdese que antes de extraer los atributos se deben haber establecido los diferentes sen-

tidos de cada nominalización (Sección 9.1.1). En la Figura 9.3 se observa que la entrada léxica del lema ‘aceptación’ consta de dos sentidos nominales. Estos dos sentidos se establecen porque tienen asociado un tipo denotativo distinto: el primero es un evento y el segundo un resultado. El sentido verbal correspondiente es el mismo en ambos sentidos (`originlink=“verb.aceptar.1”`). En la Figura 9.4 observamos solo el sentido del lema ‘golpe’ en la construcción ‘golpe de estado’. Se ha constituido como un sentido independiente porque forma parte de una construcción lexicalizada.

- **Cousin** [`“cousin={yes, no}”`]. Este atributo marca si el sentido de la nominalización se deriva morfológicamente de un verbo (`cousin=“no”`, en las Figuras 9.3 y 9.4) o es una nominalización *cousin* (`“cousin=yes”`). Recuérdese que por nominalización *cousin* se entiende aquellas nominalizaciones que solo tienen una relación semántica con los verbos o aquellas nominalizaciones en las que la relación morfológica es de sustantivo a verbo (Subsección 1.1.1).
- **Denotación** [`“denotation={result, event, underspecified, none}”`]. Este atributo hace referencia al tipo denotativo de la nominalización, es decir, indica su interpretación semántica. Los valores posibles se corresponden con los tres tipos denotativos establecidos en este trabajo, evento (`“event”`), resultado (`“result”`) y subespecificado (`“underspecified”`), y con un valor nulo (`“none”`) en el caso de los sentidos lexicalizados que se corresponden a lexías no nominales a las que no se asocia tipo denotativo. En la Figura 9.3, observamos dos sentidos, el primero de los cuales es resultativo (`denotation=“result”`) y el segundo eventivo (`denotation=“event”`). En la Figura 9.4, el sentido lexicalizado ‘golpe de estado’ constituye una lexía nominal por lo que sí se le asocia un tipo denotativo, en este caso resultativo (`denotation=“result”`).
- **Identificador** [`“id={1, 2, 3, n}”`]. Este atributo sirve para indicar el número de sentido en la entrada léxica. En la Figura 9.3 se representan dos sentidos, el primer sentido es `“id=1”` y el segundo `“id=2”`. En la Figura 9.4, el sentido lexicalizado ‘golpe de estado’ es el cuarto sentido en la entrada léxica de ‘golpe’ tal y como indica el identificador `“id=4”`.
- **Lexicalización**. [`“lexicalized={yes, no}”`]. Este atributo indica si una nominalización forma parte de una construcción lexicalizada (`lexicalized=“yes”`) como en la Figura 9.4 o no (`lexicalized=“no”`), como es el caso de los dos sentidos de ‘aceptación’ (Figura 9.3). En el primer caso, se añaden dos atributos adicionales:

- **Lema alternativo.** [“**alternativelemma**={**lemaalternativo**₁, **lemaalternativo**_n}”]. En este atributo se especifica la construcción lexicalizada completa de la que la nominalización forma parte. Por lo tanto, los valores posibles son las construcciones lexicalizadas. En la Figura 9.4, el valor para este atributo es la construcción lexicalizada “golpe de estado” (**alternativelemma**=“golpe de estado”).
- **Tipo de lexicalización** [“**lexicalizationtype**={**nominal**, **verbal**, **adjectival**, **adverbial**, **prepositional**, **conjunctive**}”]. En este atributo se declara de qué tipo de lexicalización se trata. Los valores son seis: **lexía nominal** (‘golpe de estado’), **verbal** (‘estar de acuerdo’), **adjectival** (‘al alza’), **adverbial** (‘con cuidado’), **preposicional** (‘en busca de’) y **conjuntiva** (‘en la medida que’) de acuerdo con la semejanza a las diferentes clases de palabras (sustantivo, verbo, adjetivo, adverbio, preposición y conjunción, respectivamente). solo en el caso de las lexicalizaciones nominales se asocia un tipo denotativo. En la Figura 9.4, la construcción lexicalizada ‘golpe de estado’ es una lexicalización nominal (**lexicalization type**= “nominal”) y su valor denotativo es resultado (“**denotation**= result”).
- **Lema origen** [“**originlemma**={**lema**₁, **lema**_n}”]. En este atributo se especifica el lema del verbo del cual deriva la nominalización. Por lo tanto, los posibles valores son todos los lemas verbales de los que se derive una nominalización. En la Figura 9.3, el valor para este atributo es “aceptar” en ambos sentidos (**originlemma**=“aceptar”) y en la Figura 9.4 el valor para este atributo es “golpear”.
- **Sentido verbal origen** [“**originlink**={**sentido-verbal**₁, **sentido-verbal**_n}”]. Dado que los verbos también pueden tener más de un sentido, este atributo apunta al sentido verbal concreto del que deriva la nominalización. Por lo tanto, los valores posibles son todos los sentidos verbales de los que se derive una nominalización. Recuérdese que sentidos nominales distintos ligados a una nominalización, suponen sentidos nominales diferentes. En la Figura 9.3, sin embargo, este atributo toma el mismo valor en ambos sentidos (“**originlink**=verb.aceptar.1”) ya que en este caso los sentidos se establecen por la distinción denotativa. En la Figura 9.4, se indica que este sentido nominal se deriva del primer sentido en la entrada verbal de ‘golpear’ (**originlink**=“verb.golpear.1”). Además, este atributo es también muy importante porque es el que se utiliza para establecer la relación entre las entradas nominales de AnCora-Nom y las entradas verbales de AnCora-Verb.
- **Synsets de Wordnet**, [**wordnetsynset**=“{**synset**₁, **synset**_n}”]. Finalmente, como los sustantivos del corpus AnCora-Es están anotados con synsets de

9. ANCOR-A-NOM: UN LÉXICO DE NOMINALIZACIONES DEVERBALES

WordNet , se ha incorporado también esta información al léxico. Se ha usado el *offset* para la codificación (prefijado con la versión de WordNet)⁵. En la Figura 9.3, el primer sentido de ‘aceptación’ se corresponde con dos synsets (wordnetsynset =“16:00117820+16:10039397”), mientras que el segundo sentido se relaciona con un solo synset (wordnetsynset=“16:00117820”). En la Figura 9.4, el sentido lexicalizado ‘golpe de estado’ se corresponde con un solo synset (wordnetsynset =“16:00629246”).

```
<?xml version="3.0" encoding="UTF-8"?>
<lexentry lemma="golpe" lng="es" origin="deverbal" type="noun">
  <sense alternativelemma="golpe_de_estado" cousin="no" denotation="result" id="4"
lexicalizationtype="nominal" lexicalized="yes" originlemma="golpear" originlink="verb.golpear.1"
wordnetsynset="16:00629246">
  <frame appearsinplural="no" type="default">
    <argument argument="arg0" thematicrole="agt">
      <constituent frequency="1" type="s.a"/>
      <constituent frequency="1" preposition="de" type="sp"/>
    </argument>
    <argument argument="argL">
      <constituent frequency="6" preposition="de" type="sp"/>
    </argument>
    <argument argument="argM" thematicrole="fin">
      <constituent frequency="1" preposition="en_favor_de" type="sp"/>
    </argument>
    <specifiers>
      <constituent frequency="3" posttype="indefinite" type="determiner"/>
      <constituent frequency="2" type="void"/>
      <constituent frequency="1" posttype="demonstrative" type="determiner"/>
    </specifiers>
    <reference_modifiers>
      <constituent frequency="1" type="s.a"/>
    </reference_modifiers>
    <examples>
      <example file="3LB-CAST/111_C-6.tbf.xml" nodepath="1.2.1.1" sentencenodepath="1">El
empresario fiyiano George_Speight encabeza este golpe de Estado en_favor_de la comunidad de nativos
fiyianos, como ha definido su acción . </example>
      <example file="CESS-CAST-AA/8907_20000114.tbf.xml" nodepath="0.0.1.4.3.1.1.0"
sentencenodepath="0">El ex presidente de Costa_de_Marfil Henri_Konan_Bédié , destituido el pasado
24_de_diciembre por un golpe de Estado militar, reclamó hoy en París la celebración de elecciones " libres
y transparentes " en su país antes de Junio próximo . </example>
    </examples>
  </frame>
</sense>
</lexentry>
```

Figura 9.4: Entrada léxica del sentido lexicalizado ‘golpe de estado’

⁵Se ha usado la versión WordNet 1.6 del español.

9.2.3. Atributos a nivel de marco

En esta subsección se detallan los atributos asociados a cada marco nominal. Recuérdese que antes de extraer los atributos se deben haber establecido los diferentes marcos nominales de cada sentido (Sección 9.1.1). En la Figura 9.3, cada uno de los sentidos de ‘aceptación’ solo contiene un marco del tipo “*default*”. Esto es así porque a las ocurrencias de ‘aceptación’ en el corpus les corresponde el marco verbal “*default*”, es decir, el marco menos marcado (en este caso, ‘A21.transitive-agent-patient’) del sentido 1 del verbo ‘aceptar’. En la Figura 9.4, el sentido lexicalizado ‘golpe de estado’ también contiene únicamente un marco del tipo “*default*”. Esto indica que las ocurrencias de ‘golpe de estado’ en el corpus están asociadas al marco verbal “*default*”, es decir, el marco menos marcado (‘A21.transitive-agent-patient’) del sentido 1 del verbo ‘golpear’⁶.

A continuación describimos la información declarada en los atributos a nivel de marco:

- **Tipo de marco verbal** [“**type**={**default, passive, unaccusative, benefactive, locative, resultative**}”]. Este atributo indica el marco verbal del que deriva la nominalización. En AnCora-Verb cada sentido verbal puede realizarse en uno o más marcos (el *default* o menos marcado, el pasivo, el anticausativo, el locativo, etc.) según las alternancias de diátesis en las que participe dicho sentido verbal (véase la Sección 4.1). En las entradas nominales se marcan los correspondientes marcos verbales que son los valores para el atributo “**type**”. La mayoría de las veces las nominalizaciones se derivan del marco menos marcado del verbo, como hemos visto en la Sección 9.1.1, por lo que el valor tiende a ser “*default*” como en las Figuras 9.3 y 9.4 (“**type**=*default*”).
- **Aparece en plural** (“**appearsinplural**={**yes, no**}”). Este atributo indica si alguna ocurrencia de la nominalización de un marco particular aparece en plural. Se trata de un atributo booleano. En las Figuras 9.3 y 9.4 ninguno de los marcos nominales aparece en plural, por lo que el valor es “no” (“**appearsinplural**= no”).

Como adelantamos en la Sección 9.1, a nivel de marco se declaran cuatro nodos (<**argument**>, <**reference-modifier**>, <**specifiers**> y <**examples**>) que a su vez tienen asociados distintos atributos. Los vemos a continuación.

- **Estructura argumental** (<**argument**>). En este nodo se declaran todos los argumentos del marco nominal. Los atributos asociados son:

⁶No todos los marcos verbales no marcados son ‘A21.transitive-agent-patient’, los marcos no marcados dependen del verbo concreto.

- **Posición argumental** [“**argument**={**arg0**, **arg1**, **arg2**, **arg3**, **arg4**, **argM**}”]. En este atributo se especifica la posición argumental asociada al argumento de la nominalización.
- **Papel temático** [“**argument**={**agt**, **pat**, **tem**, **cau**, **ben**, ...}”]. En este atributo se especifica el papel temático asociado al argumento de la nominalización. El conjunto de valores posibles de este atributo son los 19 papeles temáticos descritos en la Sección 3.2.

Para cada argumento se especifican los constituyentes que los pueden realizar. Esto se hace mediante los atributos siguientes:

- **Tipo de constituyente** [“**type**={**sp**, **s.a**, **determiner**, **relatiu**, **sadv**, **sn**,}”]. En este atributo se declara qué clase de constituyente puede realizar un argumento. Los constituyentes que pueden realizar todo tipo de argumentos, “nucleares” y adjuntos, son los SPs, SAs, determinantes posesivos y pronombres relativos. Los SAdvS y los SNs solo realizan argumentos adjuntos.
- **Frecuencia** [“**frequency**={**1**, **2**, **n**}”]. En este atributo se especifica el número de veces en el que el constituyente realiza al argumento en el corpus.
- **Preposición** [“**preposition**={**a**, **ante**, **bajo**, **con**, **contra**, **de**, **desde**, **durante**, **en**, **entre**, **hacia**, **hasta**, **mediante**, **para**, **por**, **pro**, **según**, **sin**, **sobre**, **tras**, **vía**}”]. Si el tipo de constituyente que realiza el argumento es un SP, entonces mediante este atributo se declara la preposición que introduce ese SP. El conjunto de valores posibles son las preposiciones del español.
- **Subtipo de determinante** [“**postype**={**possessive**}”]. El único determinante que se puede interpretar como argumento de una nominalización son los determinantes posesivos, de ahí que el único valor posible para este atributo sea “possessive”. Este atributo solo aparece si el tipo de constituyente que realiza al argumento es un determinante.

En la Figura 9.3, el sentido resultativo tiene un argumento (“arg0”) con el papel temático de agente (“agt”). Este argumento se realiza una vez (frequency=“1”) por un SP (type =“sp”) introducido por la preposición ‘de’ (“preposition=de”) y otra vez (frequency=1”) por un SA (“type=s.a”). El sentido eventivo tiene un argumento (“arg1”) con el papel temático de paciente (“pat”). Este argumento se realiza dos veces (frequency=2”) por un SP (“type =sp”) introducido por la preposición ‘de’ (“preposition=de”) y

una vez (“frequency=1”) mediante un determinante posesivo (“type=determiner”, “postype=possessive”). En la Figura 9.4, el sentido lexicalizado resultativo tiene un solo argumento (“arg0”) con el papel temático de agente (“agt”). Este argumento se realiza una vez (“frequency=1”) por un SA (“type=s.a”) y otra vez por un SP (“type=sp”) introducido por la preposición ‘de’ (“preposition=de”) (“frequency=1”).

- **Modificadores de la referencia** (<referencemodifier>). En este nodo se representan aquellos complementos nominales que no son argumentos de la nominalización pero modifican su referencia. Los atributos asociados son:
 - **Tipo de constituyente** [“type={sp, s.a, S, sadv, sn,}”]. En este atributo se declaran qué clase de constituyente puede realizar un complemento nominal no argumental. Los constituyentes que pueden realizar este tipo de complementos son los SPs, los SAs, los SAdvS, los SNs y las Ss.
 - **Frecuencia** [“frequency={1, 2, n}”]. En este atributo se especifica el número de veces en el que el constituyente realiza en el corpus al complemento nominal no argumental.
 - **Preposición** [“preposition={a, ante, bajo, con, contra, de, desde, durante, en, entre, hacia, hasta, mediante, para, por, pro, según, sin, sobre, tras, vía}”]. Si el tipo de constituyente que realiza el complemento nominal no argumental es un SP, entonces mediante este atributo se declara la preposición que introduce ese SP. El conjunto de valores posibles son las preposiciones del español.

En la Figura 9.4, el sentido lexicalizado ‘golpe de estado’ tiene un SA como modificador de referencia (“type=s.a”) realizado una vez (“frequency=1”).

- **Especificación** (<specifiers>). En este nodo se representan los elementos que ocupan la posición de especificador de los SNs cuyos núcleos son las nominalizaciones deverbales, generalmente determinates. Los atributos asociados son:
 - **Tipo de constituyente** [“type={determiner, void}”]. La posición de especificador del SN puede estar vacía, y entonces el valor asociado es “void”. Si la posición de especificador la ocupa un determinante el valor asociado es “determiner”. En este último caso, además, se añade el atributo siguiente.

- **Subtipo de determinante** [**“postype={article, demonstrative, exclamative, indefinite, interrogative, numeral, ordinal, possessive}”**]. El tipo de determinante que admite la nominalización es un criterio útil para distinguir el tipo denotativo, por lo que se incluyó este rasgo como atributo en la representación léxica de las nominalizaciones deverbales. Los posibles valores son “article” (artículo definido), “indefinite” (determinante indefinido), “demonstrative” (determinante demostrativo), “exclamative” (determinante exclamativo), “numeral” (determinante numeral), “interrogative” (determinante interrogativo), “possessive” (determinante posesivo), “ordinal” (determinante ordinal).
- **Frecuencia** (**“frequency={1, 2, n}”**). En este atributo se especifica el número de veces en el que el constituyente (un tipo de determinante o constituyente vacío) especifica en el corpus los SNs cuyos núcleos son las nominalizaciones deverbales.

En la Figura 9.3, el sentido resultativo se especifica una vez por un artículo definido (type=“determiner”, postype=“article”, frequency=“1”) y otra vez aparece sin determinante (type=“void”, frequency=“1”). El sentido eventivo se especifica dos veces por un artículo definido (type=“determiner”, postype=“article”, frequency=“2”). En la Figura 9.4, el sentido lexicalizado se especifica tres veces por un determinante indefinido (type=“determiner”, postype=“indefinite”, frequency=“3”), dos veces aparece sin determinante (type=“void”, frequency=“2”) y otra vez aparece especificado por un determinante demostrativo (type=“determiner”, postype=“demonstrative”, frequency=“1”).

- **Ejemplos** (<examples>). Cada marco nominal contiene todos los ejemplos del corpus de los cuales se ha extraído la información especificada en los atributos descritos. Para cada uno de los ejemplos <example> se declaran los siguientes atributos:
 - **Fichero** [**“file={ N}”**]. Este atributo indica el fichero del corpus en el que se encuentra el ejemplo. Por ejemplo, en la Figura 9.3 el primer ejemplo del sentido resultativo se encuentra en el fichero CESS-??CAST-??P/141_19981202.tbf.xml del corpus (file=“CESS-??CAST-??P/141_19981202.tbf.xml”).
 - **Camino a la nominalización** [**“nodepath={N}”**]. Este atributo indica el camino en el fichero hasta llegar a la nominalización correspondiente. Por ejemplo, en la Figura 9.3 en el primer ejemplo del sentido eventivo para llegar a la nominalización ‘aceptación’ se debe recorrer el siguiente camino en el fichero (nodepath=“7.4.1.1.1.3.2.1.2.0”).

- **Camino a la oración** [“**nodepath**={N}”]. Este atributo indica el camino en el fichero hasta llegar a la oración que contiene la nominalización, es decir, el número de oración empezando a partir de 0. Por ejemplo, en la Figura 9.4 en el primer ejemplo del sentido para llegar a la oración que contiene la construcción lexicalizada ‘golpe de estado’ se debe recorrer el siguiente camino en el fichero (sentence-nodepath=“1”), lo que significa que es la segunda oración del fichero porque se empieza a contar a partir de 0.

9.3. Análisis cuantitativo de los datos

Disponer de la información más importante sobre las nominalizaciones de verbales del español recopilada y estructurada en el léxico AnCora-Nom permite realizar de forma automática una serie de recuentos que nos aportan datos para el estudio de estas. A continuación destacamos algunos de estos datos.

AnCora-Nom consta de 1.655 entradas léxicas de sustantivos deverbales del español, que contienen 3.094 sentidos y 3.204 marcos. El promedio de sentidos nominales por lema es de 1,87 como se indicó anteriormente, es decir, menos de dos sentidos por lema de media. A pesar de que existen entradas nominales con más de 10 sentidos (las entradas de ‘cuenta’, ‘cambio’, ‘juego’ y ‘subida’) y que el 48 % de las entradas polisémicas tienen más de dos sentidos, más de la mitad de las entradas (53 %) son monosémicas (véase la Sección 9.1.1 para el cálculo de estos porcentajes), lo que explica este promedio.

El promedio de marcos nominales por sentidos es aún inferior (1,1), lo que significa que la mayoría de sentidos contiene un único marco nominal. De hecho, sentidos nominales con más de un marco solo hay 109 de los 3.094, lo que supone solo el 3,5 % de los sentidos. El resto de sentidos (96,5 %) tienen únicamente un marco. Entre los marcos, el más frecuente del que se derivan las nominalizaciones es el no marcado (“default”): de los 3.204 marcos, 3.066 son marcos con el valor “default” (96 %). El resto de marcos suponen porcentajes mínimos: marco de sujeto oblicuo (0,1 %), marco anticausativo (2 %), marco causativo (0,1 %), marco pasivo (1 %), marco benefactivo (0,7 %), marco resultativo (0,1 %).

Distribución de sentidos

La distribución de la muestra de sentidos nominales en función del resto de atributos, es otra muestra de los datos interesantes que se pueden obtener de este léxico. En la Tabla 9.1 se presenta la distribución de los sentidos en los distintos tipos denotativos y teniendo en cuenta si esos sentidos están lexicalizados o no.

Denotación y lexías

La denotación más frecuente entre los sentidos de las nominalizaciones es la resultativa (61 %), seguida de la denotación eventiva (20 %) y la subespecificada (16 %). Esta superioridad numérica de los sentidos resultativos no nos sorprende ya que los eventos suelen ser expresados en la mayor parte de las ocasiones por

Denotación	Lexicalizado	No lexicalizado	Total
Evento	0	631	631
Resultado	115	1.771	1.886
Subespecificado	2	490	492
Sin denotación	85	0	85
Total	202	2.892	3.094

Tabla 9.1: Distribución de los sentidos nominales: denotación y lexicalización

construcciones verbales y las nominalizaciones, como sustantivos que son, se usan más frecuentemente para un significado resultativo concreto. Además, si nos fijamos en los sentidos nominales con más ejemplos también coincide con sentidos resultativos. En concreto, los cuatro sentidos con más ejemplos del léxico son el sentido resultativo de ‘decisión’ (229 ejemplos), el sentido resultativo de ‘acuerdo’ (también 229 ejemplos), el sentido resultativo de ‘elección’ (218 ejemplos) y el sentido resultativo de ‘trabajo’ (212 ejemplos).

El 3 % restante se corresponde con sentidos de nominalizaciones que no tienen asociado ningún tipo denotativo porque forman parte de una construcción lexicalizada que no es nominal. De hecho, este grupo de construcciones lexicalizadas representan un 42 % del total de construcciones lexicalizadas, mientras que el 68 % restante son lexicalizaciones nominales, que en su mayoría son resultativas.

En la Tabla 9.2 se presenta la distribución de los sentidos nominales teniendo en cuenta el tipo denotativo y el número de argumentos que un sentido nominal tiene. Parece claro que los sentidos resultativos son los que pueden aparecer sin argumentos con mayor frecuencia (un 26 % de las veces) si los comparamos con sentidos eventivos (8 %) y subespecificados (4 %). De hecho, la mayoría de las veces los sentidos eventivos (53 %) y subespecificados (63 %) aparecen con al menos un argumento, en contraste con los sentidos resultativos (32 %). También si el número de argumentos asciende a dos, los sentidos eventivos y subespecificados son los más frecuentes (27 % y 23 % respectivamente) en comparación con los resultativos (18 %). Sin embargo, y aunque resulte chocante, son los sentidos resultativos los que aparecen un mayor número de veces con tres o más argumentos (24 %), mientras que los eventivos y los subespecificados solo tienen este número de argumentos en un 12 % y 10 % de las ocasiones, respectivamente. Este hecho se explica porque las nominalizaciones resultativas aparecen con más argumentos adjuntos (argMs), que pueden asociarse con distintos papeles temáticos (manera, tiempo, lugar) en el mismo SN. Sin embargo, los sentidos eventivos y subespecificados tienden a aparecer con argumentos “nucleares” (arg0, arg1...) y estos

Denotación y argumentos

aparecen una vez en un SN.

Argumentos	Evento	Resultado	Subespecificado
0 argumentos	48	499	22
1 argumento	336	603	312
2 argumentos	168	340	111
3 o más argumentos	79	444	47
Total	631	1.886	492

Tabla 9.2: Distribución de los sentidos nominales: denotación y número de argumentos

Denotación y determinantes

En la Tabla 9.3 se muestra la distribución de sentidos nominales teniendo en cuenta el tipo denotativo y los tipos de constituyentes (determinantes) que ocupan la posición de especificador de los SNs de núcleo de verbal. Téngase en cuenta que cada sentido nominal puede tener más de un tipo de especificador. En esta tabla parece confirmarse que los sentidos resultativos admiten una gama más amplia de especificadores que los eventivos y subespecificados. Los sentidos resultativos de las nominalizaciones se han especificado con todo tipo de determinantes excepto con determinantes ordinales (primero, segundo, etc.) y también pueden aparecer sin especificar. Los sentidos eventivos, en cambio, aparecen sin especificar o bien solo con determinantes definidos, posesivos, indefinidos y demostrativos. Sin embargo, los sentidos eventivos que aparecen con un determinante indefinido son muy pocos (16 del total de 631 sentidos eventivos, un 2,5 %) y aún menos los del determinante demostrativo (16 del total de 631 sentidos eventivos, un 0,5 %), sobre todo si los comparamos con el determinante definido (71 %) y el posesivo (26 %). Esto indica que el determinante indefinido y el demostrativo no son característicos del tipo denotativo eventivo. Los sentidos subespecificados, por su parte, también ocurren con una gama más reducida que los resultativos: no aparecen ni con determinantes exclamativos ni interrogativos. Como en los eventivos, de todos los determinantes que admiten (definido, indefinido, demostrativo, posesivo y numeral) existen algunos (indefinido, demostrativo y numeral) que no parecen ser característicos de esta clase ya que especifican a sentidos subespecificados en muy pocas ocasiones: el determinante indefinido solo especifica a un 6 % de los sentidos subespecificados, el determinante demostrativo a un 1 % y el numeral a un 0,2 %.

Otro dato que puede ser interesante en cuanto a la distribución de sentidos es en qué medida estos pertenecen al tipo *cousin*. De los 3.094 sentidos solo 98, es decir, un 3 % son *cousin*, lo que significa que la mayor parte de los sentidos de las

9. ANCORANOM: UN LÉXICO DE NOMINALIZACIONES DEVERBALES

Determinantes	Evento	Resultado	Subespecificado
Definido	451	1.305	388
Indefinido	16	885	28
Demostrativo	3	310	5
Posesivo	166	388	69
Numeral	0	92	1
Ordinal	0	0	0
Exclamativo	0	1	0
Interrogativo	0	11	0
Sin determinante	267	1.287	181

Tabla 9.3: Distribución de los sentidos nominales: denotación y tipo de determinante

nominalizaciones deverbales se derivan morfológicamente de verbos.

Además de la distribución de sentidos, otra información importante que se puede obtener mediante el análisis cuantitativo de los datos de AnCora-Nom es la realización de la estructura argumental de las nominalizaciones. ¿Qué tipo de argumento es el más frecuente? ¿Mediante qué constituyentes se realiza? En la Tabla 9.4 se presentan los distintos argumentos (posiciones argumentales) relacionados con los constituyentes que los pueden realizar. En las filas se presentan cada uno de los constituyentes que pueden realizar argumentos, tanto nucleares como adjuntos, y en las columnas se detallan cada uno de los argumentos. La columna final recoge el total de argumentos realizado por un constituyente. La última fila recoge el número total de cada tipo de argumento. Recuérdese que un argumento se puede realizar por más de un constituyente, por lo que la suma de los diferentes constituyentes que realizan un determinado argumento no coincide con el número total de cada tipo de argumento.

Realización argumental

Argumentos y constituyentes

De la Tabla 9.4 se desprende que el argumento más realizado en el dominio nominal es el arg1 (2.077), seguido de distintos argMs (1.530), arg0 (775) y arg2 (345). Estos números confirman que los arg0s (es decir, el argumento que realiza prototípicamente al agente) apenas se realizan en el dominio nominal, seguramente porque este argumento está implícito en el contexto, y que el arg1 (es decir, el argumento paciente o tema) resulta el más necesario ya que es el que completa el significado de la nominalización ('la construcción' vs. 'la construcción de la casa'). Entre los constituyentes, el SP es el constituyente que más veces realiza

Constituyentes	arg0	arg1	arg2	arg3	arg4	argM	Total
SP	709	2.138	394	27	52	1.558	4.878
SA	220	340	45	3	2	164	774
Posesivo	334	295	4	2	0	2	647
P. Relativo	10	32	2	0	0	0	44
SN	3	20	6	0	0	56	85
SAdv	0	3	5	0	3	33	44
Total	775	2.077	345	25	45	1.530	4.797

Tabla 9.4: Distribución de los distintos tipos de argumentos según el tipo de constituyente

todas las posiciones argumentales y le sigue a mucha distancia el SA. El resto de constituyentes no realizan todas las posiciones argumentales: el determinante posesivo realiza en mayor medida el arg0 y el arg1, la realización de arg2, arg3 y argMs mediante este constituyente es meramente testimonial; al pronombre relativo le ocurre lo mismo, realiza mayormente el arg0 y el arg1 (aunque con menos frecuencia que el determinante posesivo) y el resto de argumentos son testimoniales. Los SNs y SAdv realizan principalmente argumentos adjuntos, si bien en el caso del SN existe un importante número que realiza también arg1. Creemos que esto puede explicarse por SNs como ‘la alianza BBVA-Teléfonica’, en el que el SN complemento se interpreta como arg1 ‘El BBVA y Telefónica se alían’ en la oración ‘La alianza BBVA-Teléfonica cumple las normas de libre mercado’.

Todos los datos cuantitativos que hemos visto y muchos otros se pueden extraer de AnCora-Nom⁷, un recurso muy útil para quien esté interesado en las nominalizaciones deverbales del español.

9.4. Conclusiones

AnCora-Nom se ha obtenido automáticamente a partir de la información anotada en el corpus AnCora-Es. La ventaja de la metodología automática es la reducción de costes y tiempo para la creación de recursos. En este caso, hemos visto que con la anotación del corpus, que ha sido validado manualmente, se garantiza la consistencia y calidad tanto del corpus como del léxico obtenido a partir de aquél. Sin embargo, tiene un claro inconveniente: en principio, no todos los ar-

⁷Este léxico se puede consultar en la página web: <http://clic.ub.edu/corpus/ancoranom.es>.

gumentos nominales posibles están representados en el léxico sino solo aquellos que aparecen anotados en el corpus AnCora-Es. Por lo tanto, hay argumentos nominales que no quedan representados. Con la intención de paliar este problema, creímos necesario tratar de anotar los argumentos implícitos de los sustantivos nominales, es decir, aquellos argumentos que aparecen en el contexto textual de la nominalización pero que no se encuentran en el SN del cual la nominalización es núcleo. Esto nos permitirá completar la representación léxica de las nominalizaciones deverbales en AnCora-Nom. En el capítulo de conclusiones, presentamos una primera aproximación a los argumentos implícitos de los sustantivos deverbales del español, la principal línea futura de este trabajo.

★ ★ ★

CAPÍTULO 10

CONCLUSIONS AND FURTHER WORK

* Este capítulo está redactado en inglés porque así se requiere para obtener la mención europea en el título de Doctor.

In this chapter we bring together the results obtained in this work, we highlight our contributions to the research topic and we discuss our plans for immediate and future work. The first section in this chapter is devoted to the contributions to the research community generated by this work (Section 10.1). The second section details the lines of future work that have been derived from the research we have presented (Section 10.2.2).

10.1. Contributions

In this thesis we contribute to the semantic analysis of texts focusing on Spanish deverbal nominalizations. The thesis is specifically centered on the semantic denotation that characterize deverbal nominalizations and on their argument structure. Our contributions are summarized in the following list:

- A set of criteria to distinguish between event and result nominalizations in Spanish.
- A linguistic study of the syntactic realization of the argument structure of nominalizations.
- The building of the ADN-Classifier. A tool for automatically classifying deverbal nominalizations into denotation types.

-
- The implementation of RHN. A set of heuristic rules that allows for the automatic annotation of argument structure of nominalizations in the AnCora-Es corpus.
 - Annotation guidelines for the manual validation of argument structure in the AnCora-Es corpus.
 - Annotation guidelines for the manual validation of denotation in the AnCora-Es corpus.
 - Adaptation of the AnCora-Pipe tool in order to annotate deverbal nominalizations.
 - Enrichment of the AnCora-Es corpus with the semantic annotation of the denotation and argument structure of deverbal nominalizations.
 - Creation of the AnCora-Nom lexicon.
 - Primary annotation guidelines for implicit arguments.

These contributions can be classified in three axes: 1) linguistics findings about deverbal nominalizations (Section 10.1.1); 2) tools to deal with deverbal nominalizations (Section 10.1.2); and 3) resources developed that include the representation of deverbal nominalizations (Section 10.1.3). Next, we present these contributions.

10.1.1. Linguistic Findings

The linguistic findings arising from this work are concerned with the semantic denotation and argument structure of deverbal nominalizations, the two main topics we have dealt with in this thesis. First, we describe the characteristics of the argument structure in the nominal domain (Section 10.1.1.1). Next, we concentrate on the characterization of deverbal nominalizations according to their denotation and on the identification of the most useful criteria to distinguish between these denotation types (Section 10.1.1.2).

10.1.1.1. Argument Structure of deverbal nominalizations: linguistic findings

The initial assumption about the argument structure of deverbal nominalizations in this thesis is that they inherit the argument structure of the base verb, which seems to be confirmed since the RHN rule set, which mostly relies on the semantic information contained in the AnCora-Verb lexicon, achieves a global

performance of 77 % F1 without taking into account the RefMod label. Nevertheless, not all the specific linguistic observations we made during the empirical study of the argument structure (Chapter 3) that underpin the different heuristic rules created (Chapter 4) are validated to the same degree. Next we detail which rules performed better and, therefore, which linguistic assumptions are corroborated. In terms of the general rules, it can be said that they perform satisfactorily. The rules for the detection of RefMod complements achieve an F1 of 94 % (Table 4.11), which means that the hypothesis in the literature positing that non relational APs, AdvPs, NPs and Ss¹ cannot be arguments inside a nominalized NP (Badia, 2002), (Meyers, 2007) and (Picallo, 1999) is largely confirmed. However, in the case of NPs we claim that those containing a named entity “location” or “date” will be a locative (argM-loc) or temporal (argM-tmp) adjunct argument, respectively. This is partially confirmed for argM-loc assignation in NPs (50 % F1) and strongly confirmed in argM-tmp assignation in NPs (71 % F1) (Table 4.11). The rules that take into account the type of preposition introducing a PP also perform well (See Table 4.12), corroborating that certain prepositions point to specific argM tags.

The analysis of the performance of the specific rules, those that take into account the information from the base verb, shows that there is no fixed order in the realization of nominal arguments that corresponds to verbal arguments. In PPs and APs, arg1 does not always appear as the first complement and arg0 (for verbal class A) and arg2 (for class B and C) do not always appear as the second complement. However, the inverse order –arg1 associated to the second complement– would not have achieved better results. The cases analyzed from the corpus show that the order of nominalized constituents is freer than that of verb arguments and, to a certain extent, depends on the context. For the same reason, a specific syntactic-semantic pattern would not be accurate enough since the context is what provides enough information to associate the arg1 to the first or second nominalized constituent in a nominalized NP. Moreover, this freer order is also motivated by a greater degree of optionality. In fact, we observed that arg0 is an optional argument that is almost never syntactically realized in nominalizations, which is an interesting linguistic finding. Specifically for PPs, it is worth noting that governed prepositions are not always shared between the nominal and the verbal complements: arg1-∅ and arg2-∅ tags prototypically correspond to verbal PP complements. The rules for their assignation exhibit high precision, meaning that they are correct most of the time when the preposition is shared between the verbal and the nominal complements. However, the low recall indicates that there

¹In this chapter, the acronyms used along the thesis are translated to English. AP stands for adjectival phrase (SA in Spanish); AdvP for adverbial phrase (SAdv in Spanish); NP for nominal phrase (SN in Spanish); PP for prepositional phrase (SP in Spanish); Grel for genitive relative pronoun (GRel in Spanish); and S for Sentence (OSub in Spanish).

is a large number of arg1- \emptyset and arg2- \emptyset nominal PP complements that cannot be detected because the preposition is not shared between the verbal and the nominal complements. In the case of APs, we found that some relational adjectives (45 %) are non-argumental, thus calling into question our initial hypothesis based on Picallo (1999); Grimshaw (1990) proposals that these types of constituent are argumental. What emerges from the analysis of these cases is that relational adjectives are subject to the phenomenon of lexical co-occurrence, that is, they are annotated as argumental or non-argumental depending on the noun they are complementing. Regarding possessive determiners, our initial hypothesis -shared for English by Gurevich and Waterman (2009)- that this type of constituents are mostly interpreted as arguments corresponding to verbal subjects is largely confirmed. Its automatic assignment achieves an F1 of 82 %. Regarding GRel, the sample of this constituent type within nominalized NPs appearing in the corpus is too small for meaningful interpretation (only 28 occurrences). Finally, the default rule assigning the argM tag to the third or fourth AP or PP is a good choice for the latter constituent but not for the former, which is mostly corrected as Ref-Mod. This confirms that in Spanish, PPs are more likely to be argM arguments (adjuncts) of nominalizations while APs tend to modify the nominalization.

10.1.1.2. The denotation of deverbal nominalizations: linguistic findings

From this thesis a set of criteria to distinguish between denotation types has been derived, particularly during the empirical study carried out (Chapter 5). On the one hand, we analyzed which of the criteria considered in the linguistic literature, mostly devoted to the English language, were relevant for Spanish. From the empirical study, we concluded that not all the criteria posited for English seem to apply to Spanish. Among the evaluated criteria, the most relevant for distinguishing between event and result nominalizations are: 1) the semantic class of the verb from which the noun is derived; 2) its pluralization capacity; 3) its determiner types; 4) the preposition introducing an agentive complement; and 5) the obligatory presence of an internal argument. These features are represented as attributes in the nominal lexical entries of the AnCora-Nom lexicon and therefore, the ADN-Classifier uses them for the classification task it carries out. However, one of the problems is that in each criteria we find features for supporting result nominalizations but not event nominalizations, which has consequences for the degree of accuracy achieved by the ADN-Classifier in the classification of the different denotation types.

On the other hand, the empirical study has also allowed us to find new clues that support denotation types. Firstly, during the linguistic analysis, we looked for some indicators that helped us to reinforce the event reading in order to compensate for the fact that the criteria from the literature basically offer features for

supporting result nominalizations. As a result, we found the paraphrase and agent criteria as well as the selectors, which have been proved very useful to human annotators for distinguishing between an event and a result reading. However, these criteria are difficult to implement automatically and therefore are not used by the ADN-Classifier. Secondly, the observation of the symbolic rules developed by the ML techniques applied in the computational analysis has given rise to new ways for helping to decide the denotation: depending on the arguments realized in the Noun Phrase (arg1, arg2, arg0, etc.) and on the constituents that realize these arguments, the denotation of the nominalization can be predicted most of the times (Recall Tables 5.8 and 5.9).

10.1.2. Tools

Two tools have been derived from this thesis: a rule-based system aiming at annotating the argument structure of deverbal nominalizations in Spanish –RHN– and an automatic system based on ML techniques for classifying deverbal nominalizations into denotation types –ADN-Classifier. Next we describe each of them.

The RHN system is made up of 107 heuristic rules whose aim is to map a nominalized constituent to its argument and thematic role using the AnCora-Verb lexicon, the AnCora-Es corpus and the list of relational adjectives. These rules incorporate linguistic knowledge from the previous empirical study (Chapter 3). The rules are organized on a decision-list basis, that is, they are tried sequentially until the first one is successfully applied. The target of the application of these rules is a nominalized NP which is constituted by a nominalization (N) and a particular CONTEXT that may be one, two or three constituents. Each rule satisfies a condition, a logical combination of predicates over N or CONTEXT, and therefore, a semantic tag is assigned. There are two types of rules: (i) fourteen general rules based on linguistic information from AnCora-Es, and (ii) ninety-three specific rules that also take into account the information in the AnCora-Verb lexicon. RHN results achieved an F1 of 77 %, thus showing that reusing the verbal information specified in existing linguistic resources is a good approach for the annotation of deverbal nominalization argument structure. Therefore, this automated process facilitates corpus annotation, which is always a time-consuming and costly process (with a time saving of 37 %).

The ADN-Classifier is the first tool that aims to automatically classify deverbal nominalizations in event, result or underspecified denotation types, and to identify whether the nominalization takes part in a lexicalized construction in Spanish. We set up a series of experiments in order to test the ADN-Classifier under different models and in different realistic scenarios achieving good results. The ADN-Classifier has helped us to quantitatively evaluate the validity of our

RHN system

ADN-Classifier

claims regarding deverbal nominalizations. An error analysis was performed and its conclusions can be used to pursue further lines of improvements. Models including features coming from the lexicon outperform those that only take into account features from the corpus. As expected, models working at the sense level outperform those working at the lemma level. When working at the lemma level, only the combination of features from both the lexicon and the corpus provides results that outperform the baseline. It is interesting to highlight that the number of features used to support result nominalizations is significantly superior to those used to strengthen event nominalizations. For each criteria we found features for supporting result nominalizations but not event nominalizations. As an outcome, the ADN-Classifer uses more features for detecting result than event nominalizations, and therefore, achieves a greater degree of accuracy in the former than in the latter.

AnCora-Pipe

In addition to these two new tools, we have adapted the already existing AnCora-Pipe tool in order to carry out the manual validation of the automatic annotation of both argument structure and denotation types in the AnCora-Es, creating a specific perspective –*Lexical Annotator for SN*.

10.1.3. Lexical resources

Two resources have been derived from this thesis. On the one hand, we have enriched the AnCora-ES corpus with the annotation of 23,431 deverbal nominalization occurrences according to their semantic denotation and their argument structure. On the other hand, we have built AnCora-Nom from scratch, representing the 1,655 nominalization types that correspond to these occurrences. Next, we will present these two resources in detail.

AnCora-Es

This thesis contributes to the enrichment of the annotation of a previously created resource, AnCora-Es. The methodology followed to annotate deverbal nominalizations in the AnCora-Es corpus consists of two different steps for the annotation of the denotation types and the argument structure. The first step was to run two independent automated processes: one for the annotation of denotation types (i.e., result, event, and underspecified) and another for the annotation of argument structure. Secondly, and also independently, we manually checked both types of information. The final outcome, the AnCora-Es corpus, is, as far as we know, the only Spanish corpus annotated with the argument structure and the denotation of deverbal nominalizations, adding to the resources developed for English by the NomBank project (Meyers et al., 2004b; Meyers, 2007). More precisely, a total of 23,431 tokens belonging to 1,655 different types of deverbal nominalizations were annotated in AnCora-Es. A corpus annotated with this information can be very useful for many NLP tasks and applications, especially for information extraction, question answering, and nominal semantic role labelling systems for

Spanish. Furthermore, such a resource can provide real evidence for the linguistic analysis of nominalizations. Our work pointed to several interesting findings regarding the interface between syntax and semantics in nominalized NPs, such as the optionality of *arg0* arguments that map to agents, the non-fixed order of nominalizations with respect to their counterparts in a verbal environment, and the change of preposition of nominal PP complements in relation to verbal PP complements.

This thesis has also resulted in the creation of a new lexical resource: AnCora-Nom, a Spanish lexicon containing 1,655 lexical entries of deverbal nominalizations. This lexicon was developed from the information encoded in the AnCora-Es corpus. It includes all the nominalizations found in the corpus with their possible denotations and argument structure combinations. AnCora-Nom is linked to the AnCora-Es corpus and to the AnCora-Verb Spanish lexicon, constituting an excellent resource for studying the argument realization of both nouns and verbs.

AnCora-Nom

10.2. Further Work

In this section we describe the lines of work derived from this thesis. First, we focus on the work that has already been started, although it was not completed for this thesis: the (automatic) annotation of implicit arguments in line with the proposals of Gerber and Chai (2010). This will complete the argument structure annotation of deverbal nominalizations in the AnCora-Es corpus. Secondly, we outline the future work regarding both the argument structure and denotation of deverbal nominalizations in Spanish.

10.2.1. Immediate work

One inconvenience of developing the lexicon AnCora-Nom automatically (Chapter 9) from the annotation in the AnCora-Es corpus is that the arguments represented in the lexicon are not necessarily all the arguments that the nominalization may have. This is due to the fact that nominal argument structure is characterized by optionality, that is, not all the arguments are realized explicitly in the NP. Therefore, we thought it was also necessary to annotate the arguments of nominalizations that are implicit, that is, arguments that are realized in the context of the nominalization and outside the NP, in the AnCora-Es corpus. The annotation of these arguments will complete the representation of the argument structure of nominalizations in the corpus, and therefore, will also allow all the arguments of the nominalizations to be represented in the AnCora-Nom lexicon.

Annotation of implicit arguments

In a time line, we are at the very beginning of this annotation process. In fact, we have only defined our concept of the implicit argument 10.2.1.1 and some

initial criteria to annotate these type of arguments 10.2.1.2.

10.2.1.1. Definition of implicit argument

We define an implicit argument as the argument which is not realized in the NP headed by the nominalization, but instead is realized in the sentence (1) or outside it (2) context.

- (1) [Las escuelas de samba de Sao Paulo]_{iarg1-pat} han conseguido [el **apoyo** [de la empresa privada] _{arg0-agt} para mejorar las fiestas de carnaval]_{NP}.
*[Schools of samba in Sao Paulo]_{iarg1-pat} got [the **support** [of private industry] _{arg0-agt} to improve Carnival celebrations]_{NP}.*
- (2) [El carnaval de Sao Paulo es feo]_{iarg1-pat}, dijo hoy [el alcalde de Río de Janeiro]_{iarg0-agt} en una conversación informal con periodistas cariocas, y encendió la polémica. [...] [Esa **opinión**]_{NP} fue respaldada por el gobernador de Río de Janeiro, quien incluso fue más allá en su crítica al comentar que el carnaval que se organiza en Sao Paulo es “más aburrido que un desfile militar”
*[The Carnival of Sao Paulo is ugly]_{iarg1-pat}, said [the mayor of Rio de Janeiro]_{iarg0-agt} in an informal conversation with Carioca journalists, and ignited the controversy. [...] [This **opinion**]_{NP} was supported by the governor of Rio de Janeiro, who went even further in his criticism when he commented that the carnival is held in Sao Paulo is “more boring than a military parade”.*

Example (1) shows the deverbal nominalization ‘apoyo’ *support* with the agent argument (‘de la empresa privada’, *of private industry*) realized inside the NP, whereas the patient argument (‘las escuelas de samba de Sao Paulo’, *schools of samba in Sao Paulo*) is realized in the same sentence but outside the NP. In (2), the nominalization ‘opinión’, *opinion*, appears without any explicit argument in the NP. However, this does not mean that it has no arguments: the agent argument (‘el alcalde de Río de Janeiro’, *the mayor of Rio de Janeiro*) as well as the patient argument (‘el carnaval de Sao Paulo es feo’, *the carnival of Sao Paulo is ugly*) are realized implicitly in the previous sentence.

Nowadays the AnCora-Es corpus is only annotated with arguments inside the NP, but as we have seen, there are arguments that are realized implicitly. The main goal is to identify this type of arguments and assign an argument position –iarg0², iarg1, etc.– and a thematic role (agent, patient, cause) to them. These arguments can be recovered if a wider discursive context is taken into account (Ruppenhofer et al., 2010) and their identification, therefore, is important to provide a deep

²The letter ‘i’ at the beginning of the argument position stands for implicit argument.

semantic representation of texts.

IARG-AnCora will be the first corpus annotated with implicit arguments in Spanish, which in turn will facilitate the enrichment of the representation of the argument structure in AnCora-Nom. At present, the only corpora with nominal implicit arguments have been developed for English and they have been used as training data for the works presented in Ruppenhofer et al. (2010) and Gerber and Chai (2010). The number of occurrences annotated are 3,073 in the former and 1,253 in the latter. Both corpora are annotated only with core arguments (no adjuncts arguments). In contrast, IARG-AnCora will have an extended coverage in two senses: on the one hand, all the implicit arguments of all deverbal nominalization occurrences in the corpus AnCora-Es (23,431) will be annotated; on the other hand, we will take into account the core arguments (arg0, arg1, arg2, arg3 and arg4) as well as the adjunct arguments (argM). IARG-AnCora will be the first corpus with this information with a high coverage available to the research community and could be used as a learning corpus for SRL nominal systems.

10.2.1.2. Criteria for the annotation of implicit arguments

Next, we summarize the criteria that we initially propose for the annotation of implicit arguments. These criteria are just some initial ideas to proceed in the annotation of implicit argument. Of course, they have to be evaluated in a inter-annotator agreement test: checking whether they are clear and useful enough to carry out this type of annotation.

The first thing to make clear is the unit to be explored for detecting implicit arguments of nominalizations. We consider the sentence where the nominalization appears (current sentence in Figure 10.1) and the sentences before (sentence -1 in Figure 10.1) or after (sentence +1 in Figure 10.1) in the document. The aim is to find constituents outside the NP that semantically represent nominalization arguments not realized within the NP. The candidates to be implicit arguments are obtained by looking at the argument structure specified in the nominal lexicon AnCora-Nom and in the verbal lexicon AnCora-Verb: those arguments represented in the verbal or nominal lexical entries and not realized in the NP are candidates to be implicit arguments. In Figure 10.1, the nominalization ‘decisión’, *decision* appears without arguments in the NP where it appears, therefore the arg0-agt (who decides), the arg1-pat (what is decided) and the different adjunct arguments are candidates to be implicit arguments. To find these implicit arguments it is necessary to look at the whole current sentence in which the nominalization appears as well as the previous (-1) and following (+1) sentences.

Secondly, we specify the constituents that can be implicit arguments. We believe that all type of constituents can be implicit arguments of nominalizations, however we require the annotated constituent to occupy the highest possible po-

into account the other constituents (mentions) that are part of that entity.

In the example of Figure 10.1, the implicit arg1-pat of ‘decisión’, *decision* (‘que revisará los permisos de pesca...’, *that will revise the fishing licenses*), is associated with an entity and, therefore, the remaining mentions of this entity are not taken into account. Note that the second mention of this entity corresponds to the same nominalization we are annotating (‘decisión’, *decision*), and we still associate this entity as the implicit arg1-pat. This is possible because the first mention of the entity is outside the NP. Likewise, if the constituent that is the implicit argument of the nominalization is below the NP headed by the nominalization, we only mark this implicit argument if this constituent is part of an entity in which the remaining mentions are indeed outside the NP.

Node name	Func/Arg/Tem/Ls	Lemma	Contents
▼ sentence			
▼ sn	subj/arg0/agt		
▶ spec			La
▶ grup.nom			Federación_Gallega_de_Baloncesto
▼ grup.verb			
v	a2	emitir	emitió
▼ sadv	cc/argM/tmp		
▶ grup.adv			hoy
▼ sn	cd/arg1/pat		
▶ spec			un
▼ grup.nom			
n	[arg0] noun.comu	comunicado	comunicado
▼ S	cn//		
▶ sp	cc/argM/loc		en el que
sn	∅ - subj/arg0/agt		
▶ grup.verb			asegura
▼ S	cd/arg1/pat		
▶ conj			que
▶ sn	subj/arg0/agt		el presidente del Colegio_Gallego_de_Árbitros , Luis_Angel_Sabariz
▶ grup.verb			dimitió
▶ sp	cc/arg2/tem		de su cargo
▶ sn	cc/argM/tmp		el pasado día_25
▶ sp	cc/argM/loc		en la reunión de la comisión delegada de la FGB
f			.

Figura 10.2: Syntactic structure of sentence (4)

- (4) La Federación Gallega de Baloncesto emitió hoy [un **comunicado** [en el que asegura que el Presidente del Colegio gallego de Árbitros, Luis Ángel Sabariz, dimitió de su cargo el pasado día 25 en la reunión de la comisión delegada de la FGB]_{S-RefMod}]_{NP}.

*The Galician Federation of Basketball today issued [a **statement** [in which they announced that the President of the Galician Association of Arbitrators, Luis Angel Sabariz, resigned the last day 25th in the meeting of the Executive Committee of the FGB]_{S-RefMod}] NP.*

In (4), whose syntactic structure is represented in Figure 10.2, it could be unders-

tood that the subordinated clause ‘que el presidente [. . .] dimitió de su cargo. . . ’ is an implicit arg1 argument. However, since this subordinated clause is syntactically embedded in the NP whose head is ‘comunicado’ (see Figure 10.2), we do not annotate it because it cannot be said that the implicit argument is outside the NP. However, if this subordinated clause was part of an entity whose other mentions were outside the NP, then we would annotate that entity as an iarg1.

If the constituent that represents the implicit argument is not associated with any entity, this constituent is immediately considered to be a new entity with a sole mention, that is, a singleton. In the case of ‘decisión’, *decision* in Figure 10.1, the implicit argument arg0-agt is represented by the PP located in the current sentence, ‘por el Ministerio de Economía’, *by the Ministry of Economy*, which is immediately considered to be a new entity.

When the constituent selected as an implicit argument is not part of an entity, this constituent must be the closest to the nominalizations. However, there are two exceptions:

1. if the implicit argument is a pronoun attached to the verb (‘arreglado’, *fix it*), we do not select this verb as an implicit argument. Therefore, we select the previous constituent that makes reference to that pronoun,
2. if we understand that the implicit argument is an apposition, it does not have to be marked. Instead, the parent NP is selected.

Finally, we are working on some precaution for assigning the thematic role to implicit arguments. In order to assign the thematic role to implicit arguments the information in the AnCora-Nom and AnCora-Verb lexicons can be consulted. However, there are some cases where a specific consign should be done. For instance, in sports predicates as ‘empate’, *tie* or ‘victoria’, *victory*, the arg2-atr would be associated to the result of the match and the argM-adv would be the opposing team.

- (5) [El Zaragoza]_{iargM-adv} empató contra el Atlético de Madrid [2-2]_{iarg2-atr} [. . .]
 [El Atlético de Madrid]_{iarg0-agt} cedió [un **empate**]_{NP}.
 [Zaragoza]_{iargM-adv} tied against Atletico Madrid [2-2]_{iarg2-atr} [...] [Atletico de Madrid]_{iarg0-agt} gave [a **draw**]_{NP}.

Once the annotation guidelines are finished, the next step is to carry out an inter-annotator test for the annotation of implicit arguments. The inter-annotator agreement test will be conducted on a sample of one hundred sentences from the AnCora-Es corpus, each sentence containing a true deverbal nominalization with at least two possible implicit arguments, that is, arguments that are in the AnCora-Nom or AnCora-Verb lexical entries and are not realized in the NP. Three Linguistics graduate students at the University of Barcelona will participate in the

test. All of them will have experience in the annotation of coreference and argument structure for nominalizations in the AnCora-Es corpus, but we still will carry out a training process on a sample of one hundred sentences from the AnCora-Es corpus with the same requirement. The inter-annotator agreement test will allow us to check if this criteria work and correct them if necessary.

10.2.2. Future work

Besides the immediate work presented, we plan to work on the improvement and transportability of the two main tools developed in this thesis, the RHN system and the ADN-Classifier, as well as on the application of the resources generated.

Regarding RHN, one line of improvement will concentrate on the detection of true argumental APs and PPs. Especially, in the case of APs, it was proved that “relational adjectives” are not always argumental, as previously thought. We want to observe in more depth the surrounding context of argumental and non-argumental APs in order to extract some interesting clues that allow us to better detect argumental APs. In this sense, future work will consist of applying these improved heuristic rules to Catalan and studying the transportability of these rules to similar Romance languages.

Regarding ADN-Classifier, two of the main sources of error found in its performance are data sparseness of some of the features (such as PP *agent*) and the fact that there are criteria at the disposal of human annotators that the ADN-Classifier is unable to detect. In order to reduce the problem of data sparseness it would be interesting to look for some linguistic generalizations of the sparse features in order to implement a backoff mechanism. Another line of future work is to analyze the criteria used by human annotators and not currently implemented either in the lexicon or in the corpus. Some additional features could be incorporated in the ADN-Classifier. Among them are path-based syntactic patterns that have been successfully applied to related tasks (See Gildea and Palmer (2002)).

We have also experimented with a meta-classifier working on the results of binary classifiers (one for each class). The global accuracy of the meta-classifier was not greater than that of the current ADN. We think, however, that a binary classifier for the *underspecified* type (the most difficult one) could result in improvements.

Other point of future work consists of analyzing to what extent the ADN-Classifier and its models are applicable to other languages. Concretely, since we have a similar corpus for Catalan (lacking deverbal nominalization information) we plan to apply the models learned for Spanish to this closely related Romance language.

Besides carrying out improvements to the tools presented and applying them to other languages, we also plan to use the resources obtained from this thesis for

NLP tasks such as PP attachment or the automatic recognition of light verbs.

Another line of future work will consist of applying the presented methodology (automatic processes and manual validations) to annotate the nominalizations in the Catalan AnCora-Ca corpus. This will form the basis for future comparative linguistic studies of Spanish and Catalan. In the future, we also intend to enlarge the AnCora-Nom lexicon with deadjectival nominalizations and relational nouns since we consider that they can also have an argument structure. We also intend to build a similar lexicon for Catalan nominalizations.

BIBLIOGRAFÍA

- Abeillé, A., Clément, L., and Kinyon, A. (2000). Building a treebank for French. In *In Proceedings of the Second International Language Resources and Evaluation (LREC'00)*, pages 87–94. European Language Resources Association (ELRA).
- Alexiadou, A. (2001). *The Functional Structure in Nominals. Nominalizations and Ergativity*. John Benjamins. Amsterdam/Philadelphia.
- Alonso, M. (2004). *Las construcciones con verbos de apoyo*. Visor Libros.
- Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.
- Aparicio, J., Taulé, M., and Martí, M. (2008). AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 797–802, Marrakech, Morocco. European Language Resources Association (ELRA).
- Aston, G. and Burnard, L. (1998). *The BNC Handbook: exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.
- Atserias, J., Rigau, G., and Villarejo, L. (2004a). Spanish WordNet 1.6: Porting the Spanish Wordnet across Princeton versions. In *In Proceedings of the Fourth International Language Resources and Evaluation (LREC'04)*, pages 1–6.
- Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. (2004b). The MEANING Multilingual Central Repository. In *Proceedings of the Second International WordNet Conference-GWC 2004*, pages 23–30.

-
- Badia, T. (2002). Els complements nominals. In Solà, J., editor, *Gramàtica del Català Contemporani*, volume 3, pages 1591–1640. Empúries. Barcelona.
- Badia, T. and Saurí, R. (2008). Developing a Generative Lexicon within HPSG. preprint.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL'98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Balvet, A., Barque, L., Condetto, M.-H., Haas, P., Huyghe, R., Marín, R., and Merlo, A. (2011). Nomage: an electronic lexicon of French deverbal nouns based on a semantically annotated corpus. In *Proceedings of the International Workshop on Lexical Resources (WoLeR) at European Summer School in Logic, Language and Informatio (ESSLLI 2011) (to appear)*.
- Balvet, A., Barque, L., and Marín, R. (2010). Building a Lexicon of French Deverbal Nouns from a Semantically Annotated Corpus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1408–1413, Valletta, Malta. European Language Resources Association (ELRA).
- Barque, L., Huyghe, R., Jugne, A., and Marín, R. (2009). Two types of deverbal activity nouns in French. In *Proceedings of the 5th International Conference on Generative Approaches to the lexicon*, pages 169–175, Pisa, Italy.
- Bertran, M., Borrega, O., Recasens, M., and Soriano, B. (2008). AnCoraPipe: A tool for multilevel annotation. *Procesamiento del Lenguaje Natural.*, 41:291–292.
- Boleda, G. (2007). *Automatic acquisition of semantic classes for adjectives*. PhD thesis, Pompeu Fabra University, Barcelona, Spain.
- Borer, H. (1997). The morphology interface: A study of autonomy. In Dressler, W. U., Prinzhorn, M., and Reunison, J. R., editors, *Advances in Morphology*, pages 5–30. Mouton de Gruyter.
- Bos, J. (2008). Wide-Coverage Semantic Analysis with Boxer. In Bos, J. and Delmonte, R., editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

- Bosque, I. and Picallo, C. (1996). Postnominal adjectives in Spanish DPs. *Journal of Linguistics*, 32, pp 349-385 doi:10.1017/S002222670001592.
- Bresnan, J., editor (1982). *The Mental Representation of Grammatical Relations*. The MIT Press.
- Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2009). FrameNet for the semantic analysis of German: Annotation, representation and automation. In Boas, H. C., editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pages 209–242. Mouton de Gruyter.
- Butnariu, C., Kim, S. N., Nakov, P., Ó Séaghdha, D., Szpakowicz, S., and Veale, T. (2009). SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105, Boulder, Colorado. Association for Computational Linguistics.
- Butnariu, C., Kim, S. N., Nakov, P., Ó Séaghdha, D., Szpakowicz, S., and Veale, T. (2010). SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 39–44, Uppsala, Sweden. Association for Computational Linguistics.
- Carmona, J., Cervell, S., Luís. Márquez, M. A. M., Lluís Padró, R. P., Horacio Rodríguez, M. T., and Turmo, J. (1998). An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the First conference on Language Resources and Evaluation (LREC'98)*, pages 923–931, Granada, Spain. European Language Resources Association (ELRA).
- Che, W., Li, Z., Hu, Y., Li, Y., Qin, B., Liu, T., and Li, S. (2008). A cascaded syntactic and semantic dependency parsing system. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL'08*, pages 238–242.
- Chklovski, T. and Mihalcea, R. (2002). Building a sense tagged corpus with open mind word expert. In *Proceedings of the ACL-02 workshop on Word Sense Disambiguation: recent successes and future directions - Volume 8, WSD '02*, pages 116–122, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.

-
- Chomsky, N. (1970). Remarks on Nominalization. In Jacobs, R. and Rosenbaum, P., editors, *Readings in English Transformational Grammar*, pages 184–221. Waltham, Mass.: Ginn and Company.
- Chomsky, N. (1981). *Lectures on Government and Binding*. Foris, Dordrecht.
- Chomsky, N. (1995). *The Minimalist Program*. Cambridge, Massachusetts: MIT Press.
- Ciaramita, M., Attardi, G., Dell’Orletta, F., and Surdeanu, M. (2008). DeSRL: A Linear-Time Semantic Role Labeling System. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, CoNLL’08, pages 258–262.
- Civit, M. and Martí, M. (2004). Building Cast3LB: A Spanish Treebank. *Research on Language and Computation*, 2(4):549–574.
- Clark, H. H. (1975). Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, TINLAP ’75, pages 169–174, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Cohen, W. W. (1995). Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.
- Copestake, A. (2007). Semantic composition with (Robust) Minimal Recursion Semantics. In *Proceedings of the Workshop on Deep Linguistic Processing*, DeepLP ’07, pages 73–80. Association for Computational Linguistics.
- Creswell, C., Beal, M. J., Chen, J., Cornell, T. L., Nilsson, L., and Srihari, R. K. (2006). Automatically extracting nominal mentions of events with a bootstrapped probabilistic classifier. In *Proceedings of the Computational Linguistics/Association for Computational Linguistics on Main conference poster sessions*, COLING-ACL ’06, pages 168–175, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Decadt, B., Hoste, V., Daelemans, W., and Bosch, A. V. D. (2004). GAMBL, Genetic Algorithm Optimization of Memory-Based WSD. In *Proceedings of the Association of Computational Linguistics /SIGLEX Senseval-3*, pages 108–112, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Demonte, V. (1989). *Teoría Sintáctica: de las Estructura a la Rección*. Madrid: Síntesis.

- Dowty, D. (1979). *Word Meaning and Montague Grammar*. Reidel, Dordrecht.
- Eberle, K. (2004). Flat underspecified representation and its meaning for a fragment of German. Technical report, Universität Stuttgart.
- Eberle, K., Faasz, G., and Heid, U. (2009). Corpus-based identification and disambiguation of reading indicators in German nominalizations. In *Online Proceedings of the 5th Corpus Linguistics Conference*.
- Eberle, K., Faasz, G., and Ulrich, H. (2011). Approximating the disambiguation of some German nominalizations by use of weak structural, lexical and corpus information. *Procesamiento del Lenguaje Natural.*, 46:67–75.
- Eberle, K., Heid, U., Kountz, M., and Eckart, K. (2008). A Tool for Corpus Analysis using partial Disambiguation and Bootstrapping of the Lexicon. In *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008*.
- Erk, K. and Padó, S. (2006). Shalmaneser: a flexible toolbox for semantic role assignment. In *Proceedings of the Fifth International Language Resources and Evaluation (LREC'06)*, pages 527–532. European Language Resources Association (ELRA).
- Fellbaum, C. (1998). *An electronic lexical database*. The Mit Press.
- Fillmore, C. J. (1968). The case for case. In Bach, E. W. and Harms, R. T., editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart & Winston, New York.
- Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.
- Fillmore, C. J. and Baker, C. F. (2001). Frame semantics for text understanding. In *Proceedings of the Workshop on WordNet and Other Lexical Resources, NAACL*, Pittsburgh, Pennsylvania. Association for Computational Linguistics.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. John Wiley.
- Gerber, M., Chai, J., and Meyers, A. (2009). The role of implicit argumentation in nominal srl. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 146–154, Boulder, Colorado. Association for Computational Linguistics.

-
- Gerber, M. and Chai, J. Y. (2010). Beyond NomBank: a study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1583–1592, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Gildea, D. and Palmer, M. (2002). The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 239–246, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Girju, R., Giuglea, A.-M., Olteanu, M., Fortu, O., Bolohan, O., and Moldovan, D. (2004). Support vector machines applied to the classification of semantic relations in nominalized noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS '04, pages 68–75, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Girju, R., Moldovan, D., Tatu, M., and Antohe, D. (2005). On the semantics of noun compounds. *Computer, Speech and Language*, 19(4):479–496.
- Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P. D., and Yuret, D. (2009). Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105–121.
- Grimshaw, J. (1990). *Argument Structure*. The Mit Press. Cambridge, Massachusetts.
- Gràcia i Solé, L. (1995). *Morfologia Lèxica: L'herència de l'estructura argumental*. Universitat de València.
- Gurevich, O., Richard, C., Holloway King, T., and De Paiva, V. (2006). Deverbal Nouns in Knowledge Representation. In *Proceedings of Florida Artificial Intelligence Research Society Conference*, pages 670–675.
- Gurevich, O. and Waterman, S. (2009). Mapping Verbal Argument Preferences to Deverbals. In *Proceedings of the 2009 IEEE International Conference on Semantic Computing*, pages 17–24.
- Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., ÓSéaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2009). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99, Boulder, Colorado. Association for Computational Linguistics.

- Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., ÓSéaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.
- Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 268–275, Morristown, NJ, USA. Association for Computational Linguistics.
- Hoeg Muller, H. (2010). Annotation of Morphology and NP structure in the Copenhagen Dependency Treebanks (CDT). In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories held at Tartu, Estonia*, pages 151–163.
- Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90 % solution. In *Proceedings of Human Language Technologies - North American chapter Association Computational Linguistics (HLT-NAACL'06)*, pages 57–60.
- Hull, R. D. and Gomez, F. (2000). Semantic interpretation of deverbal nominalizations. *Natural Language Engineering*, 6(2):139–161.
- Jezek, E. and Melloni, C. (2009). Complex types in the (morphologically) complex lexicon. In *Proceedings of the 5th International Conference on Generative Approaches to the lexicon*, pages 59–67, Pisa, Italy.
- Johansson, R. and Nugues, P. (2008). Dependency-based Syntactic–Semantic Analysis with PropBank and NomBank. In *Proceedings of the Twelfth Conference on Natural Language Learning, CoNLL'08*, pages 183–187, Manchester, United Kingdom.
- Kipper, K., Dang, H. T., Schuler, W., and Palmer, M. (2000). Building a class-based verb lexicon using TAGs. In *Proceedings of the Fifth International Workshop on Tree Adjoining Grammars and Related Formalisms*, Paris, France.
- Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2006). Extending VerbNet with novel verb classes. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1027–1032, Genova, Italy.
- Lapata, M. (2002). The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

-
- Levi, J. N. (1978). *The Syntax and Semantics of Complex Nominals*. New York, San Francisco, London: Academic Press Inc.
- Loper, E., Yi, S., and Palmer, M. (2007). Combining lexical resources: Mapping between PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, pages 1–12.
- Macleod, C., Grishman, R., Meyers, A., Barrett, L., and Reeves., R. (1998). NOM-LEX: A Lexicon of Nominalizations. In *Proceedings of EURALEX'98*, pages 187–193.
- Madnani, N. and Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.
- Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.
- Martí i Girbau, N. (2002). El SN: els noms. In Solà, J., editor, *Gramàtica del Català Contemporani*, volume 3, pages 1281–1335. Empúries. Barcelona.
- McLachlan, G., Do, K., and Ambroise, C. (2004). *Analyzing microarray gene expression data*. Wiley.
- Mechura, M. (2008). *Selectional Preferences, Corpora and Ontologies*. PhD thesis, Trinity College, University of Dublin, Dublin, Ireland.
- Meinschaefer, J. (2005). Deverbal nouns in Spanish. *Linguae et linguaggio*, IV, 2: 215-228.
- Mel'cuk, I. (1981). Meaning-Text Models: A Recent Trend in Soviet Linguistics. *Annual Review of Antropology*, 10, 27-62.
- Mel'cuk, I., Arbatchewsky-Jumaire, N., Elnitsky, L., and Iordanskaja, L. (1984). *Dictionnaire explicatif et combinatoire du français contemporain*. Presses de l'Université de Montréal.
- Meyers, A. (2007). Annotation Guidelines for NomBank Noun Argument Structure for PropBank. Technical report, University of New York.
- Meyers, A., Reeves, R., Macleod, C., Szekeley, R., Zielinska, V., and Young, B. (2004a). The Cross-Breeding of Dictionaries. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal.

- Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004b). Annotating Noun Argument Structure for NomBank. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal.
- Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., and Girju, R. (2004). Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS '04, pages 60–67, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Mooney, R. J. (2007). "learning for semantic parsing". In *Computational Linguistics and Intelligent Text Processing: Proceedings of the 8th International Conference (CICLing 2007) (invited paper)*, pages 311–324. Springer, Berlin, Germany.
- Márquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue. *Computational Linguistics*, 34(2):145–159.
- Nakov, P. I. (2007). *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. PhD thesis, EECS Department, University of California, Berkeley.
- Nunes, M. L. (1993). Argument Linking in English Derived Nominals. In Valin, R. D. V., editor, *Advances in Role Reference Grammar*, pages 375–432. John Benjamins. Amsterdam/Philadelphia.
- Ohara, K. (2009). Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru. In Boas, H. C., editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pages 163–182. Mouton de Gruyter.
- Padó, S., Pennacchiotti, M., and Sporleder, C. (2008). Semantic role assignment for event nominalisations by leveraging verbal data. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, COLING '08*, pages 665–672, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Palmer, M. (2009). SemLink: Combining English Lexical Resources. In *Proceedings of the Generative Lexicon Conference, GenLex-09*, pages 19–25.
- Palmer, M., Gildea, D., and Xue, N. (2010). *Semantic Role Labeling. Synthesis on Human Languages Technologies*. Morgan and Claypool Publishers.

-
- Palmer, M., Kingsbury, P., and Gildea, D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):76–105.
- Palmer, M. S., Dahl, D. A., Schiffman, R. J., Hirschman, L., Linebarger, M., and Dowding, J. (1986). Recovering Implicit information. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 10–19, New York, New York, USA. Association for Computational Linguistics.
- Peris, A. (2011). AnCora-Nom: Guía de anotación para la Estructura Argumental de los sustantivos deverbales. Working paper 3: TEXT-MESS 2.0 (Text-Knowledge 2.0). Technical report, University of Barcelona.
- Peris, A. and Taulé, M. (2009). Evaluación de los criterios lingüísticos para la distinción evento y resultado en los sustantivos deverbales. In *Proceedings of the 1st International Conference on Corpus Linguistics*, pages 596–611, Murcia, España.
- Peris, A. and Taulé, M. (2011a). AnCora-Nom: A Spanish Lexicon of Deverbal Nominalizations. *Procesamiento del Lenguaje Natural.*, 46:11–19.
- Peris, A. and Taulé, M. (2011b). Annotating the argument structure of deverbal nominalizations in Spanish. doi: 10.1007/s10579-011-9172-x. *Language Resources and Evaluation*.
- Peris, A., Taulé, M., Boleda, G., and Rodríguez, H. (2010a). ADN-Classifier: Automatically Assigning Denotation Types to Nominalizations. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1422–1428, Valletta, Malta.
- Peris, A., Taulé, M., and Rodríguez, H. (2009). Hacia un sistema de clasificación automática de sustantivos deverbales. *Procesamiento del Lenguaje Natural.*, 43:23–31.
- Peris, A., Taulé, M., and Rodríguez, H. (2010b). Semantic Annotation of Deverbal Nominalizations in the Spanish AnCora Corpus. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, pages 187–198, Tartu, Estonia.
- Peris, A., Taulé, M., and Rodríguez, H. (2012). Empirical methods for the study of denotation in nominalizations in Spanish. *Computational Linguistics*. To appear.

- Philpot, A., Hovy, E., and Pantel, P. (2005). The omega ontology. In *Proceedings of IJCNLP Workshop on Ontologies and Lexical Resources (OntoLex-05)*, volume 280, pages 59–66.
- Picallo, C. (1999). La estructura del Sintagma Nominal: las nominalizaciones y otros sustantivos con complementos argumentales. In Bosque, I. and Demonte, V., editors, *Gramática Descriptiva de la Lengua Española*, volume 1, pages 363–393. Espasa Calpe. Madrid.
- Pollard, C. and Sag, I. A. (1987). *Information-based syntax and semantics: Vol. 1: fundamentals*. Center for the Study of Language and Information, Stanford, CA, USA.
- Pollard, C. and Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.
- Pradhan, S. S., Loper, E., Dligach, D., and Palmer, M. (2007). Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval '07*, pages 87–92, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pustejovsky, J. (1995). *The Generative Lexicon*. The Mit Press. Cambridge, Massachusetts.
- Pustejovsky, J., Knippen, R., Littman, J., and Saurí, R. (2005). Temporal and event information in natural language text. *Language Resources and Evaluation*, 39:123–164.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.
- Rainer, F. (1999). La derivación Adjetival. In Bosque, I. and Demonte, V., editors, *Gramática Descriptiva de la Lengua Española*, volume 3, pages 4595–4642. Espasa Calpe. Madrid.
- Rappaport, M. (1983). On the Nature of derived Nominals. In Levin, B., Rappaport, M., and Zaenen, A., editors, *Papers in Lexical Functional Grammar*, pages 113–142. Waltham, Mass.: Ginn and Company.
- Real Academia de la Lengua Española (2012). *Diccionario de la Lengua Española*. Versión electrónica.
- Recasens, M. (2010). *Coreference: Theory, Annotation, Resolution and Evaluation*. PhD thesis, University of Barcelona, Barcelona, Spain.

-
- Recasens, M. and Martí, M. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44:315–345.
- Recasens, M., Martí, M. A., and Taulé, M. (2007). Text as Scene: Discourse Deixis and Bridging Relations. *Revista de la Asociación Española para el Procesamiento del Lenguaje Natural*, 39:205–212.
- Recasens, M. and Vila, M. (2010). On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2006). FrameNet II: Extended Theory and Practice. Technical report, International Computer Science Institute.
- Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., and Palmer, M. (2009). Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 106–111, Boulder, Colorado. Association for Computational Linguistics.
- Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., and Palmer, M. (2010). SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 296–299, Uppsala, Sweden. Association for Computational Linguistics.
- Santiago, R. and Bustos, E. (1999). La derivación Nominal. In Bosque, I. and Demonte, V., editors, *Gramática Descriptiva de la Lengua Española*, volume 3, pages 4505–4594. Espasa Calpe. Madrid.
- Saurí, R. and Pustejovsky, J. (2009). Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.
- Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.
- Sebastián, N., Martí, M. A., Carreiras, M., and Cuetos, F. (2000). *LEXESP: Léxico Informatizado del Español*. Barcelona. Ediciones de la Universitat de Barcelona.
- Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.
- Spencer, A. and Zaretskaya, M. (1999). The Essex Database of Russian Verbs and their Nominalizations. Technical report, University of Essex.

- Subirats, C. (2009). Spanish FrameNet: A frame semantic analysis of the Spanish Lexicon. In Boas, H. C., editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pages 135–162. Mouton de Gruyter.
- Surdeanu, M., Johansson, R., Meyers, A., Márquez, L., and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, CoNLL '08, pages 159–177, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Taulé, M., Martí, M., and Recasens, M. (2008). AnCorà: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 96–101, Marrakech, Morocco. European Language Resources Association (ELRA).
- Vendler, Z. (1967). *Linguistics in Philosophy*. Cornell University Press.
- Vila, M., Martí, M. A., and Rodríguez, H. (2011). Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach. *Procesamiento del Lenguaje Natural*, 46:83–90.
- Vossen, P., editor (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.
- Vossen, P. and Fellbaum, C. (2009). Universals and idiosyncrasies in multilingual WordNets. In Boas, H. C., editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pages 319–345. Mouton de Gruyter.
- Vázquez, G., Fernández, A., and Martí, M. A. (2000). *Clasificación verbal. Alternancias de diátesis*. Quaderns de Sintagma, 3, Edicions de la Universitat de Lleida.
- Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edition.
- Xue, N. (2006). Semantic role labeling of nominalized predicates in Chinese. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 431–438, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yi, S., Loper, E., and Palmer, M. (2007). Can Semantic Roles Generalize Across Genres? In *Proceedings of Human Language Technologies -North American chapter Association Computational Linguistics (HLT-NAACL' 07)*, pages 548–555.

Zhao, H. and Kit, C. (2008). Parsing Syntactic and Semantic Dependencies with Two Single-Stage Maximum Entropy Models. In *Proceedings of the Twelfth Conference on Natural Language Learning, CoNLL'08*, pages 203–207, Manchester, United Kingdom.

Zubizarreta, M. L. (1987). *Levels of Representation in the Lexicon and in the Syntax*. Foris. Dordrecht.

Apéndices

APÉNDICE A

LISTA DE ADJETIVOS RELACIONALES

Aberzale, accidental, activista, adicional, administrativo, aéreo, aeroespacial, aeroportuario, agrario, agroalimentario, agropecuario, alavesista, albanés, alemán, alimentario, ambiental, american, anal, ancestral, andalucista, andaluz, anímico, anticastrista, anticonstitucional, antidroga, antifranquista, antimilitarista, antimisil, antisindical, antiterrorista, antropológico, anual, aragonés, arbitral, argumental, armamentista, armeniocanadiense, arqueológico, arquitectónico, artesanal, artificial, artístico, asistencial, asiático, atómico, audiovisual, austriaco, automático, automovilista, automovilístico, autonómico, autoritario, balcánico, bancario, barcelonista, biomédico, brasileño, británico, burocrático, callejero, catalán, canario, capitalista, castellano-Leonés, castrista, catalanista, cerebral, cervical, chabacano, chileno, chino, chino-japonés, ciclista, científico, cinematográfico, circense, ciudadano, cívico, civil, climático, clínico, colectivo, colonial, comercial, comunista, comunitario, conservador, constitucional, constructivo, continental, contractual, contraterrorista, coruñés, cónico, criminal, crítico, cubano, cultural, danés, defensivo, delictivo, democrático, deportivista, deportivo, diario, divino, doctrinal, doctrinario, documental, doméstico, domiciliario, dominicano, farmacéutico, familiar, Fantasmagórico, fascista, federal, federativo, femenino, ferroviario, filipino, financiero, finlandés, fiscal, físico, fitosanitario, flemático, floral, fluvial, foral, forestal, fotográfico, francés, franquista, fraternal, fraudulento, funcionarial, futbolístico, gallego, gastrointestinal, generacional, genético, geográfico, geopolítico, geotécnico, gijonés, ginecológico, golpista, grancanario, gravitacional, gremial, gripal, gubernamental, hepático, hispánico, histórico, holandés homosexual, hotelero, humano, humanitario, humorístico, idiomático, imperial, impresionista, industrial, informático, inglés, inmobiliario, institucional, instrumental, interbancario, interclasista, intercontinental, intergubernamental, interministerial, interna-

cional, iraquí, islámico, israelí, italiano, izquierdista, judicial, jurídico, jurídico-político, jurisdiccional, laboral, laborista, leonés, libanés lingüístico, literario, luxemburgués, madridista, mallorquin, marroquí, maternal, medioambiental, mediterráneo, mecánico, mensual, metálico, metalúrgico, milanista, milanés, ministerial, mobiliario, monetario, multicultural, multilateral, multinacional, multipartidista, mundial, municipal, musical, nacional, nacionalista, naval, neogaullista, neuronal, nominal, norteamericano, nórdico, numeral, occidental, ocupacional, oficial, operacional, orbital, oriental, papal, paraempresarial, paragubernamental, parlamentario, parroquial, partidista, patronal, peatonal, pélvico, peneuvista, pericial, periodístico, peronista, plurinacional, pluripartidista, policial, político, portuario, portugués, posicional, postcomunista, postelectoral, potencial, preelectoral, presidencial, presupuestario, procesal, protocolario, provincial, psicológico, publicitario, quebequés, racial, radiofónico, rafaelista, rayista, recreativista, renal, residencial, revolucionario, ritual, salarial, sanitario, semanal, semestral, semilaboral, sindical, social, socialcomunista, socialista, soviético, sueco, suroccidental, táctico, teatral, técnico, tecnológico, telefónico, temporal, territorial, terrorista, testifical, thailandés, tradicional, transcultural, tropical, turinés, turístico, ugetista, universitario, urbano, vacacional, valencianista, vasco, vecinal, vegetal, venezolano, verbal, vespertino, vigués, virginal, vital, vocal, zapatista, y zarista.

APÉNDICE B

LISTA DE PUBLICACIONES RELACIONADAS CON LA TESIS

- Peris, Aina, Mariona Taulé y Horacio Rodríguez (2012). ‘Empirical methods for the study of denotation in nominalizations in Spanish’. *Computational Linguistics*. (aceptado, pendiente de publicación).
- Peris, Aina y Mariona Taulé (2011). ‘Annotating the argument structure of deverbal nominalizations in Spanish’. *Language Resources and Evaluation*. DOI: 10.1007/s10579-011-9172-x
- Peris, Aina y Mariona Taulé (2011). ‘AnCora-Nom: A Spanish lexicon of deverbal nominalizations’. *Procesamiento del Lenguaje Natural*, nº46, pp. 11-18. Jaén, España.
- Peris, Aina, Mariona Taulé y Horacio Rodríguez (2010). ‘Semantic Annotation of Deverbal Nominalizations in the Spanish corpus AnCora’. *Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pp. 187-198, University of Tartu, Estonia.
- Peris, Aina, Mariona Taulé, Gemma Boleda y Horacio Rodríguez (2010). ‘ADN-Classifier: Automatically assigning denotation types to nominalizations’. *Proceedings of the 7th International Conference on Language Resources and Evaluation*. La Valleta, Malta.
- Peris, Aina, Mariona Taulé y Horacio Rodríguez (2009). ‘Hacia un sistema de clasificación automática de sustantivos deverbales’. *Procesamiento del Lenguaje Natural*, nº 43, pp 23–31. Jaén, España.

- Peris, Aina y Mariona Taulé (2009). 'Evaluación de los criterios lingüísticos para la distinción evento y resultado en los sustantivos deverbales. *Proceedings of the 1st International Conference on Corpus Linguistics (CILC-09)*. Murcia, España.