Empirical Methods for the Study of Denotation in Nominalizations in Spanish

Aina Peris*
University of Barcelona

Mariona Taulé*
University of Barcelona

Horacio Rodríguez**
Technical University of Catalonia

This article deals with deverbal nominalizations in Spanish; concretely, we focus on the denotative distinction between event and result nominalizations. The goals of this work is twofold: first, to detect the most relevant features for this denotative distinction; and, second, to build an automatic classification system of deverbal nominalizations according to their denotation. We have based our study on theoretical hypotheses dealing with this semantic distinction and we have analyzed them empirically by means of Machine Learning techniques which are the basis of the ADN-Classifier. This is the first tool that aims to automatically classify deverbal nominalizations in event, result, or underspecified denotation types in Spanish. The ADN-Classifier has helped us to quantitatively evaluate the validity of our claims regarding deverbal nominalizations. We set up a series of experiments in order to test the ADN-Classifier with different models and in different realistic scenarios depending on the knowledge resources and natural language processors available. The ADN-Classifier achieved good results (87.20% accuracy).

1. Introduction

The last few years have seen an increasing amount of work in the semantic treatment of unrestricted text, such as Minimal Recursive Semantics in Lingo/LKB (Copestake 2007), Frame Semantics in Shalmaneser (Erk and Padó 2006), Discourse Representation Structures in Boxer (Bos 2008), and the automatic learning of Semantic Grammars (Mooney 2007), but we are still a long way from representing the full meaning of texts when not restricted to narrow domains. Many Natural Language Processing (NLP) applications such as Question Answering, Information Extraction, Machine Reading and high-quality Machine Translation or Summarization systems, and many NLP intermediate level tasks such as Textual Entailment, Paraphrase Detection, or Word

* CLiC, Centre de Llenguatge i Computació/University of Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona. E-mail: aina.peris@ub.edu; mtaule@ub.edu.
** TALP Research Center - Technical University of Catalonia, Jordi Girona Salgado 1-3, 08034 Barcelona. E-mail: horacio@lsi.upc.edu.

Submission received: 22 July 2011; revised submission received: 27 December 2011; accepted: 1 February 2012.

Sense Disambiguation (WSD), have almost reached their practical upper bounds and it is difficult to move forward without using a serious semantic representation of the text under consideration. Given the limitations and the difficulties in obtaining an in-depth semantic representation of texts as a whole, most efforts have been focused on partial semantic representation using less expressive semantic formalisms, such as those that come under the umbrella of Description Logic variants, or on discarding the whole semantic interpretation task in order to focus on smaller (and easier) subtasks. This is the case, for instance, in Semantic Role Labeling (SRL) systems, which indicate the semantic relations that hold between a predicate and its associated participants and properties, the relations of which are drawn from a pre-specified list of possible semantic roles for that predicate or class of predicates. See Márquez et al. (2008) and Palmer, Gildea, and Xue (2010) for recent surveys. Closely related to SRL is the task of learning Selectional Restrictions for a predicate, for example, the kind of semantic class each argument of the predicate must belong to (Mechura 2008). In this case a predefined set of semantic classes must also be used to perform the classification task. WordNet (Fellbaum 1998), VerbNet (Kipper et al. 2006), PropBank (Palmer, Kingsbury, and Gildea 2005), FrameNet (Ruppenhofer et al. 2006), and OntoNotes (Hovy et al. 2006) are resources frequently used for this purpose. Most of these efforts are verb-centered and reduce role labeling to the roles played by entities around a predicate instantiated as a verb. At a finer level, there is the task of WSD, for example, assigning the most appropriate sense to each lexical unit of the text from a predefined lexical-semantic resource. Once again a catalogue of classes has to be used as a range for the assignment.[1] In this case as well, despite its excessive finer granularity, WordNet is the most widely used reference. See Navigli (2009) for a recent survey.

In this line of research, there has recently been a growing interest in going beyond verb-centered hypotheses to tackle the computational treatment of **deverbal nominalizations** (nouns derived from verbs), in order to move forward to the full comprehension of texts. Deverbal nominalizations are lexical units that contain rich semantic information equivalent to a clausal structure. Many recent studies have focused on the detection of semantic relations between pairs of nominals that belong to different Nominal Phrases (NPs), such as Task 4 of SemEval 2007 (Girju et al. 2009) and Task 8 of SemEval 2010 (Hendrickx et al. 2010), or between nominals taking part in noun compound constructions. In the latter case, they take into account a predefined set of semantic relations (Girju et al. 2005) or use verb paraphrases with prepositions (Task 9 of SemEval 2010 [Butnariu et al. 2010a, 2010b]). Although these works include nominalizations, they are not strictly focused on them but cover all type of nouns. Actually, most of the work studying only deverbal nominalizations is focused on their argument structure: Some authors focus on the detection of arguments within the NP headed by the nominalization (Hull and Gomez 2000; Lapata 2002; Gurevich et al. 2006; Padó, Pennacchiotti, and Sporleder 2008; Surdeanu et al. 2008; Gurevich and Waterman 2009), whereas others center their attention on detecting the implicit arguments of the nominalizations which are outside the NP (Gerber and Chai 2010; Ruppenhofer et al. 2010). Among the former group, there are different approaches to the problem: Lapata (2002) and Gurevich and Waterman (2009) use probabilistic models, Hull and Gomez (2000) and Gurevich et al. (2006) develop heuristic rules, Padó, Pennacchiotti, and

1 Some approaches simply discriminate between different senses for a case without assigning it to a predefined specific class, however. Clustering techniques rather than classification are used in these approaches.

Sporleder (2008) work with an unsupervised SRL system, and in Surdeanu et al. (2008) the work presented uses supervised SRL systems. The kind of argument annotated is also different in these works: Although only two, more syntactic labels (subj [subject] and obj [object]), are used to annotate the arguments in Lapata (2002), Gurevich et al. (2006), and Gurevich and Waterman (2009), Padó, Pennacchiotti, and Sporleder (2008) use FrameNet labels and Surdeanu et al. (2008) use NomBank (Meyers, Reeves, and Macleod 2004)[2] labels. The interpretation of nominalizations is crucial because they are common in texts and an important amount of information is represented within them. In the AnCora-ES corpus (Taulé, Martí, and Recasens 2008), for instance, the semantic information is mostly coded in verbs (56,590 verbal occurrences) but a significant number of deverbal nominalizations (23,431 occurrences) also encode rich semantic information.

Most of the work on this topic sets out from the denotative distinction between nominalizations referring to an *event*, those that express an action or a process, and nominalizations referring to a *result*, those expressing the outcome of an action or process. From a theoretical point of view, it is stated that this denotative distinction may have repercussions on the argument-taking capability of deverbal nominalizations. Despite being aware of this distinction, computational approaches focus on *event* nominalizations, not taking into account the *result* ones or, more frequently, without characterizing the difference. For instance, SRL systems are mostly applied to *event* nominalizations (Pradhan et al. 2004; Erk and Padó 2006; Liu and Ng 2007). *Result* nominalizations are more frequent than the *event* types, however, at least in Spanish (1,845 *event* occurrences in contrast to 20,037 *result* occurrences in AnCora-ES). In the present work, we hypothesize that *result* nominalizations, like *event* nominalizations, can take arguments; therefore, discarding *result* nominalizations would imply a loss of semantic information, equally relevant to text representation. In this article, we focus our interest on this denotative distinction. Concretely, we aim to determine the relevant linguistic information required to classify deverbal nominalizations as *event* or *result* types in Spanish. In order to achieve this goal, we have built an automatic classifier of deverbal nominalizations—the ADN-Classifier—for Spanish, aimed at identifying the semantic denotation of these nominal predicates (Peris et al. 2010). The ADN-Classifier is a tool that takes into account different levels of linguistic information depending on its availability, such as senses, lemmas, or syntactic and semantic information coded in the verbal and nominal lexicons (AnCora-Verb [Aparicio, Taulé, and Martí 2008] and AnCora-Nom [Peris and Taulé 2011]) or in the AnCora-ES corpus.

Therefore, this article contributes to the semantic analysis of texts focusing on Spanish deverbal nominalizations, although the proposal presented could be extended to other Romance languages. We base our study on theoretical hypotheses that we analyze empirically, and as a result we have developed three new resources: 1) the ADN-Classifier, the first tool that allows for the automatic classification of deverbal nouns as *event* or *result* nominalizations; 2) the AnCora-ES corpus enriched with the annotation of deverbal nominalizations according to their semantic denotation, the only Spanish corpus that incorporates this information; and 3) AnCora-Nom, a lexicon of deverbal nominalizations containing information about denotation types and argument structure.

The ADN-Classifier can be used independently in NLP tasks, such as Coreference Resolution and Paraphrase Detection (Recasens and Vila 2010). For Coreference

2 In the work of Hull and Gomez (2000) it is not stated explicitly which set of arguments are used, although from their examples we infer that they are semantic roles such as those of VerbNet.

Resolution tasks it would be useful to have the nominalizations classified into denotations in order to detect coreference types. For instance, if a nominalization has a verbal antecedent (anchor) and its denotation is of the *event* type, an identity coreference relation could be established between them (Example (1)). If the nominalization is of the *result* type, however, the relation established between verb and noun would be a bridging coreference relation (Example (2)) (Clark 1975; Recasens, Martí, and Taulé 2007).

(1) En Francia los precios **cayeron** un 0,1% en septiembre. **La caída**<event> ha provocado que la inflación quedara en el 2,2%.
'In France prices **fell** by 0.1 % in September. **The fall**<event> caused inflation to remain at 2.2 %.'

(2) La imprenta **se inventó** en 1440. **El invento**<result> permitió difundir las ideas y conocimientos con eficacia.
'The printing press **was invented** in 1440. **This invention**<result> allowed for ideas and knowledge to be spread efficiently.'

As for Paraphrase Detection (Androutsopoulos and Malakasiotis 2010; Madnani and Dorr 2010), *event* nouns (but not *result* nouns) are paraphrases for full sentences, so this type of information can also be useful for this task. For instance, the sentence in Example (3) and the NP headed by an *event* nominalization in Example (4) are typically considered to be paraphrases.

(3) **Se ha ampliado** el capital de la empresa en un 20%.
'The company's capital **has been increased** by 20%.'

(4) **La ampliación**<event> del capital de la empresa en un 20%.
'**The increase**<event> of company's capital by 20%.'

(5) Se han vendido **muchas traducciones**<result> de su último libro.
'**Many translations**<result> of his latest book have been sold.'

(6) Se han vendido **muchos libros traducidos** de su último título.
'**Many translated editions** of his latest book have been sold.'

If the nominalization, however, has a *result* interpretation as in Example (5)—*traducciones*, 'translations' refers to the concrete object, that is, the book translated—it is impossible to have a paraphrase with a clausal structure. This is due to the fact that *result* nominalizations can denote an object whereas verbs cannot denote objects. In fact, *result* nominalizations can only be paraphrases of other NPs denoting objects (Example (6)).

The AnCora-ES corpus enriched with denotative information could be used as training and test data for WSD systems. The work presented in this article also provides an additional insight into the linguistic question underlying it: the characterization of deverbal nominalizations according to their denotation and the identification of the most useful criteria for distinguishing between these denotation types.

The remainder of this article is organized as follows. Section 2 summarizes the theoretical approaches to the semantic denotation with which we deal here. Section 3 describes the methodology used in this work. Section 4 presents the empirical linguistic study in which the initial model is established; in Section 5 the different knowledge resources used are presented, paying special attention to the nominal lexicon, AnCora-Nom. In Section 6, the ADN-Classifier is presented and in Section 7 the different

experiments conducted are described and their results are reported. Section 8 reviews related work and, finally, our conclusions are drawn in Section 9.

2. Theoretical Background

In the linguistics literature related to nominalizations, one of the most studied and controversial topics is the denotative distinction between *event* and *result* nominalizations. By *event* nominalization we mean those nouns that denote an action or process in the same way that a verb does. In other words, as their verbal counterparts, *event* nominalizations have the aspectual property of dynamicity (Example (7)). In contrast, a *result* nominalization refers to the state (Example (9)) or the concrete or abstract object resulting from the event (Example (8)). Both types of *result* nominalizations (states and objects) lack the aspectual property of dynamicity.

(7) El proyecto americano consiste en la **adaptación**<event> de la novela Paper Boy.
'The American project is the **adaptation**<event> of the Paper Boy novel.'

(8) Esta **adaptación**<result> cinematográfica ha recibido buenas críticas.
'This film **adaptation**<result> has received good reviews.'

(9) Reforzó la **tendencia**<result> al alza del Euro de los últimos días.
'The upward **trend**<result> of the euro has been reinforced in recent days.'

In Example (7), the noun *adaptación* ('adaptation') denotes an *event* because it expresses an action in the same way that a verb does (it is equivalent to *El proyecto americano consiste en adaptar la novela Paper Boy*, 'The American project consists of adapting the Paper Boy novel'). The *event* interpretation is characterized as dynamic because it implies a change from 'not being adapted' to 'being adapted.' In contrast, in Example (8) the same nominalization is understood as a *result* because it denotes a specific object that is the outcome of the action of adapting a creative work into a film. In Example (9), the *result* interpretation is due to the fact that the verb base of *tendencia*, 'trend,' denotes a state, so the noun inherits the property of stativity (non-dynamicity) and does not imply any change.

In this sense, our notions of *event* and *result* are equivalent to the *complex-event* and *result* nominalizations, respectively, in the terminology of Grimshaw (1990)[3] or the terms *process* and *result* used in Pustejovsky (1995)[4] and Alexiadou (2001). Although the *event vs. result* distinction we make is widespread, it is true that *event* and *result* nominalizations can also be represented in a more fine-grained way. For instance, Eberle, Faasz, and Heid (2009) distinguish between *events* (*messung*, 'measurement'), *states* (*teilung*, 'division'), and *objects* (*lieferung*, 'furnished material') in German nominalizations, and Balvet, Barque, and Marín (2010) propose a typology of French nominalizations that contemplates four aspectual types: *states* (*admiration*, 'admiration'), *durative events* (*opération*, 'operation'), *punctual events* (*explosion*, 'explosion'), and *objects* (*bâtiment*, 'building'). For English, Levi (1978) identifies four types of nominalizations: *actions* (*parental refusal*), which are equivalent to the event notion; *agents* (*financial*

3 She distinguishes a third denotative type, *simple event* nouns like *trip* in *That trip took two weeks*, but we discard them because, although expressing an *event*, they are not derived from verbs; in this research we focus on deverbal nominalizations.

4 This author characterizes nominalizations as *dot-objects* that include both *process* and *result* meanings.

analyst), which denote the agent of an action and are characterized by using a different set of suffixes; *products* (*human error*), which denote the result of an action; and *patients* (*student's inventions*), which denote the patient object of the action. Also for English, Nunes (1993) defines five types of nominalizations: *process* nouns, that name the action or process denoted by the base verb (*The documents' destruction by the North*); *result* nouns, that denote a new creation resulting from the base verb (*The invention was put on display*); *accumulated-action* nouns, that name the sum total of a verb activity (*The attack was unexpected*); *experiential-state* nouns, nominalized stative verbs or nominalizations of a state brought about by a particular verb (*Sam's interest in maths*); and *experiential-state results*, the result counterpart of the previous class (*Sam has many interests*).

The authors working on this topic mainly differ on two issues. On the one hand, they do not agree on how to consider (and therefore, how to represent) these two denotations: as two senses of the same lexical unit (Pustejovsky 1995; Badia 2002; Alonso 2004) or as two different lexical units (Grimshaw 1990; Picallo 1999; Alexiadou 2001). Regarding this denotative distinction, several linguistic criteria have been proposed in order to identify each of these denotations, mostly for English, although there are some proposals for Spanish (Picallo 1999), French, Greek, Russian, and Polish (Alexiadou 2001) (see Table 2 in Section 4.1). On the other hand, authors differ on the argument-taking capacity of deverbal nominalizations: Some linguists maintain that only *event* deverbal nominalizations can take arguments (Zubizarreta 1987; Grimshaw 1990), whereas others consider that both *event* and *result* nominalizations can take arguments (Pustejovsky 1995; Picallo 1999; Alexiadou 2001).

Authors who conceptualize *event* and *result* nominalizations as different lexical units justify this in different ways. Grimshaw (1990) considers that only *complex-event* nominals legitimate an argument structure, and that constitutes the main difference with respect to *result* nominalizations which, according to her, lack argument structure. In Alexiadou (2001) and Picallo (1999), the idea that *event* and *result* nominalizations are different lexical units is justified by the different functional projection of these two types of nominalization and by word-formation at different levels of the language,[5] respectively. In contrast to Grimshaw, however, they state that both types of nominalizations can select arguments. In the words of Alexiadou (2001, page 69): "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." Picallo also believes that complements of *result* nominalizations are arguments when argumental relationships can be established between them and the nominal head.

In contrast, those who consider both denotations as senses of the same lexical unit maintain that nominalizations are *underspecified* lexical units (Pustejovsky 1995), units in which a disjunction of meaning is present (Alonso 2004), or simply lexical units with different senses (Badia 2002). Specifically, Pustejovsky accounts for the *event-result* ambiguity in nominalizations by means of an underspecified lexical representation called *dot-object*, arguing that *event-result* nominalizations are cases of complementary polysemy: "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, page 31). In relation to argument-taking capacity, both types of nominalizations are argumental because the *dot-object* representation has an argument structure in which the nominal arguments are specified.

5 In Picallo (1999) it is stated that event nominalizations are created in the syntax whereas result nominalizations are created in the lexicon; therefore they have different derivation processes.

Alonso (2004) argues that these nominalizations present a disjunction of meaning because they can update an *event* and a *result* reading in the same sentence without the understanding of the sentence being affected. For instance, *declaración* ('declaration') in Example (10)[6] can be interpreted as an *event* and as a *result* at the same time: Only *event* nominalizations can be said to have an initial moment and only *result* nominalizations can be said to have five pages. These two senses both originate in the same lexical unit, which includes both, and the context provides both meanings.

(10) La **declaración**<event/result> que el juez tomó al testigo, que comenzó a las once, ocupa cinco folios.
'The **statement**<event/result> that the judge took from the witness, which began at eleven, takes up five pages.'

Regarding argument-taking capacity, Alonso maintains that all nominalizations taking part in a support-verb construction can select arguments. So, if a *result* nominalization is part of a support-verb construction, it will also have argument structure. Following these authors, we also consider that *result* nominalizations can take arguments.

Within a computational framework, there are different models that represent nominalizations; not all of them take into account the *event* and *result* distinction, however. For instance, in NomBank (Meyers, Reeves, and Macleod 2004) this distinction is completely ignored and the authors focus on argument structure. In contrast, Spencer and Zaretskaya (1999) label each nominal sense with one of the Grimshaw semantic categories (i.e., *complex event*, *simple event*, and *result*). Their database contains information about 7,000 Russian verbs and their 5,000 corresponding nominalizations, distinguishing between the verbal entries that nominalize the whole *event* while preserving the verbal argument structure, and those that denote a concrete or abstract *result* of the verb. The Nomage project (Balvet, Barque, and Marín 2010) annotates French deverbal nominalizations in the French TreeBank (Abeillé, Clément, and Kinyon 2000) with one of the four classes proposed in their work (i.e., *states*, *durative events*, *punctual events*, and *objects*). In the middle ground between these two positions, we find the representation models proposed in WordNet (Fellbaum 1998), FrameNet (Baker, Fillmore, and Lowe 1998; Ruppenhofer et al. 2006), and Ontonotes (Hovy et al. 2006). WordNet, possibly due to its extremely fine granularity, usually includes among the senses corresponding to deverbal nouns one paraphrased as "the acting of verb x" (our *event* nominalization) and another paraphrased to something similar to "the thing X-verb+ed" (our *result* nominalization). FrameNet distinguishes between deverbal nominalizations defined as the *action* or *process* of X verb and nominalizations defined as *entities*. Concerning nouns, Ontonotes is interested in the disambiguation of noun senses in order to create an ontology. Within deverbal nouns the authors distinguish between nominalization senses that truly inherit the verb meaning and deverbal noun senses whose denotation is not directly related to the verb meaning.[7] That is, they distinguish between two different types of deverbal nouns but this distinction is not akin to the *event-result* one. The distinctions established in WordNet and FrameNet are more similar to the one proposed in OntoNotes.

To sum up, the *event vs. result* distinction in deverbal nominalizations has received much attention in linguistics literature. It seems to be less relevant in the computational

6 This example is taken from Alonso (2004).
7 For instance, *building* in *The building was made mostly of concrete and glass*.

framework, in contrast, although some computational models do represent a semantic distinction that is similar to the one we are analyzing (see Section 8).

3. Methodology

The aim of the current work is twofold: first, to detect the most relevant features for the denotative distinction between *event* and *result* nominalizations; and, second, to build

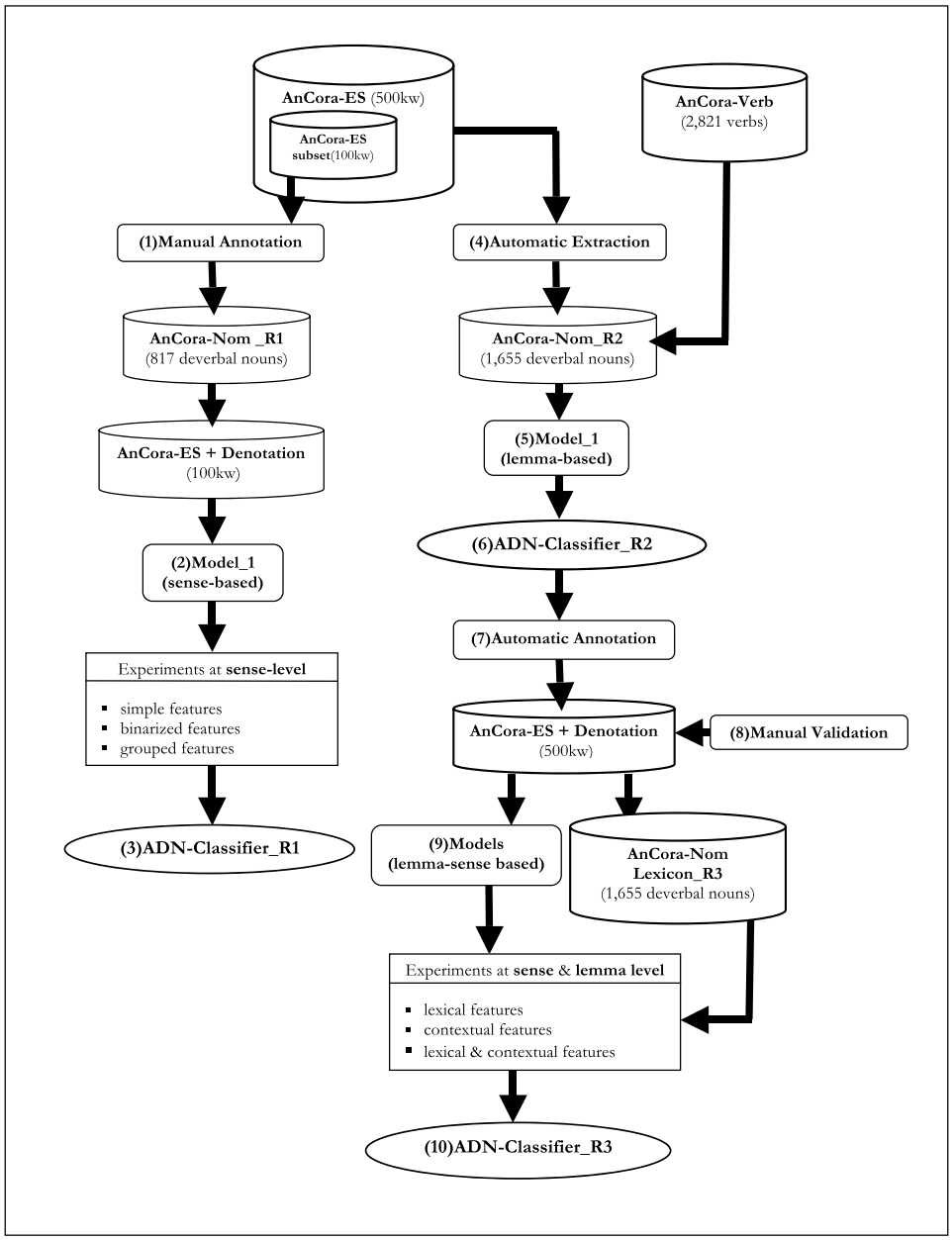

Figure 1
Scheme of the methodology followed.

an automatic classification system for deverbal nominalizations according to their denotation. In Figure 1 the overall methodology for carrying out this work is schematized.

In order to determine the most relevant features, the first step was to carry out a linguistic study of deverbal nominalizations (see Section 4). This study consisted of the application of the linguistic criteria stated in the literature to a reduced set of nominalizations corresponding to the occurrences extracted from a 100,000 word (henceforth 100kw) subset of the AnCora-ES corpus. As an outcome, we manually obtained (see step 1 in Figure 1) a lexicon for this deverbal nominalization set, AnCora-Nom-R1 (Peris, Taulé, and Rodríguez 2009), which allowed us to annotate the corresponding deverbal occurrences in the corpus subset. The nominalization classifying model (see step 2 in Figure 1) underlying these two initial resources was tested by empirical methods. This first model of classification is based on sense distinctions—that is, the extraction of features was performed at sense level and the examples for learning (both positive and negative) corresponded to different senses in the lexicon; thus, we will refer to it as the sense-based model. We set up a series of experiments based on Machine Learning (ML) techniques in order to evaluate the consistency of the data annotated in AnCora-Nom-R1, to analyze the relevance of the attributes used in the lexical-semantic representation and to infer new useful features to distinguish between *event* and *result* nominalizations. As well as experimenting with AnCora-Nom-R1 simple features, we experimented with the binarization[8] and grouping[9] of several of them to address sparseness problems. Furthermore, with these experiments the foundations of the automatic classification system, the ADN-Classifier-R1, were laid (see step 3 in Figure 1). See Section 6 for a description of the ADN-Classifier.

Once the consistency of the annotated data was corroborated and the most relevant features were established, we focused on the building of the ADN-Classifier, the second goal of the research presented in this article. In order to develop the final version of the ADN-Classifier (R3) we increased the corpus data used for learning. Therefore, we needed to annotate the denotation types in the whole AnCora-ES. Because this involves an increase in the number of nominalization occurrences to annotate (23,431 occurrences in contrast to 3,077), we carried out this annotation automatically. To do so, we modified the (sense-based) model of classification underlying the ADN-Classifier-R1 (see step 2 in Figure 1) creating a new model able to carry out the more realistic task of classifying the occurrences in the whole AnCora-ES corpus.

This new model (see step 5 in Figure 1) uses the following as knowledge resources: 1) the AnCora-Verb lexicon to obtain the features from the verbs related to nominalizations; 2) the complete AnCora-ES corpus (500kw); and 3) AnCora-Nom-R2, an extended lexicon of nominalizations without denotation types obtained automatically (see step 4 in Figure 1). This lexicon contains a total of 1,655 lexical entries, corresponding to the 1,655 nominalization types in the whole AnCora-ES. Because we annotate the occurrences in the AnCora-ES corpus, however, we reduce our dependence on the lexical source AnCora-Nom-R2 (see Section 5.3), removing the sense-specific information from it and taking into account only the information shared by all the senses of one lemma, while maintaining the features extracted from the corpus. In this sense, we adapted the sense-based model developed in ADN-Classifier-R1 (see step 2 in Figure 1) to obtain a new classification model that is based on lemmas (and not senses) (see step 5 in

8 Binarizing a *k*-value categorial feature means transforming it into *k* binary features.

9 Combining several simple features into a complex feature using a combination (for instance, a logical OR).

Figure 1). This new model was used for the automatic annotation of the AnCora-ES corpus with denotation information (see step 7 in Figure 1). Afterwards, in order to evaluate the performance of the developed model, the corpus annotation was manually validated (Peris, Taulé, and Rodríguez 2010) (see step 8 in Figure 1). This manual process also guarantees the quality of the corpus annotation, leading to the final version of the AnCora-Nom lexicon (R3) containing denotation type information.

At that moment, we were able to build the final version of the ADN-Classifier (R3) (see step 10 in Figure 1). In order to do so, we set up a series of experiments leading to the development of new sense- and lemma-based models using the resources already built (AnCora-ES with denotation information and AnCora-Nom-R3), that is, models learned with more instances. We also replicated the experiments at sense and lemma level with the subset of 100kw from the already enriched AnCora-ES and the subset of 817 lexical entries from the AnCora-Nom-R3. Specifically, we carried out a set of new ML experiments using the simple and binarized features from the nominal and verbal lexicons as well as additional features obtained from the AnCora-ES corpus (the so-called contextual features). For the evaluation of the different sense- and lemma-based models derived from this set of experiments (see step 9 in Figure 1), tenfold cross-validation was used with the pre-created resources. These models give rise to the final version of the ADN-Classifier (R3) (see step 10 in Figure 1). See Section 6 for a description of the ADN-Classifier.

4. Previous Linguistic Study

The aim of the corpus-based linguistic study conducted was twofold. First, we wanted to determine which of the criteria stated in the literature were the most relevant linguistic features to establish the distinction between *event* and *result* denotations in Spanish. Secondly, we wanted to find other features that could be used to reinforce the semantic distinction we are dealing with.

In order to do this, we selected a sample of 817 Spanish deverbal nominalizations corresponding to 3,077 occurrences. These nominalizations were obtained semiautomatically from a 100kw subset of the AnCora-ES corpus. In this selection we took into account a predefined list of ten suffixes (*-a, -aje, -ión/-ción/-sión/-ón, -da/-do, -dura/-ura, -e, -ido, -miento/-mento, -ncia/-nza, -o/-eo*– [Santiago and Bustos 1999]) that contribute to an *event* or *result* meaning and which take verbs as a basis for the derivation process.

This sample corresponds to the original 3LB corpus (Civit and Martí 2004), that can be considered to be a preliminary version of AnCora-ES. The set of 817 nominalizations consists of those occurrences in this sub-corpus. Despite coming from different sources, the 100kw corpus and the full 500kw corpus are comparable as is shown in Table 1.

In Table 1 we present some metrics for describing the whole AnCora-ES corpus and its 100kw subset. We present the metrics we have used in three rows: degree

Table 1
Descriptive content of AnCora-ES and its 100Kw subset. In each cell the values corresponding to the subset and the whole corpus are present.

	Min	Max	Mean	Standard Deviation
sense/lemma	1/1	13/13	2.19/1.86	1.54/1.31
examples/lemma	1/1	239/255	19.99/14.15	30.76/26.44
length sentences	4/4	149/149	39.14/39.51	10.69/12.08

of polysemy (number of senses per lemma), the number of examples (sentences) per lemma in the corpus, and the average length of sentences per lemma. We depict the minimum and maximum values, the mean and the standard deviation for each metric. The values in the figure seem reasonable. The only anomalous figures correspond to the extremely high values of the standard deviation of the number of examples metric. It is due to the highly biased shape of the curve towards small values. In fact, most of the lemmas have only one example (121 for the 100kw sample) and the number of lemmas having values over the mean are very few. As can be seen in Table 1, there are no notable differences in the values corresponding to the whole set and the subset. Additionally we computed the ratio of lemmas containing only one example.

4.1 Analyzing Features from the Literature

In order to determine which criteria stated in the literature were the most relevant, we selected seven criteria that satisfy two conditions: first, they are some of the most widely used by other authors, and second, it is possible to search for them in the AnCora-ES corpus without suffering data sparseness. These criteria and the authors who propose them are shown schematically in Table 2.

The seven criteria were analyzed by contrasting them with the behavior of the 817 Spanish deverbal nominalizations in the AnCora-ES corpus. Concretely, two graduate students in linguistics classified the 3,077 occurrences independently into *event* or *result* class. After this first annotation, we discussed the results with them and reached an agreement in those cases in which the denotation type assigned was not the same. It should be noted that the aim of this analysis was to encourage reflection on the suitability of these seven criteria and on finding new clues to interpret the nominalizations semantically.

During the classification procedure, we observed that these two denotations did not allow us to account for all the data in the corpus. On the one hand, it is not always possible to distinguish between *event* and *result*, because the linguistic context (the sentence) is not always informative enough. We label such cases *underspecified* types, resulting finally in three possible denotation values. On the other hand, we noticed that some nominalizations can take part in a lexicalized construction, thus, we added the attribute <*lexicalized*>. In such cases, we distinguish between six types of lexicalization according to their similarity to different word classes: nominal (e.g., *síndrome de abstinencia*, 'withdrawal symptoms'), verbal (e.g., *estar de acuerdo*, 'to be in agreement'), adjectival (e.g., *al alza*, 'rising'), adverbial (e.g., *con cuidado*, 'with care'), prepositional (e.g., *en busca de*, 'in search of'), and conjunctive (e.g., *en la medida que*,

Table 2
Linguistic criteria for distinguishing between *result* and *event* nominalizations from different authors.

Criteria	Grimshaw'90	Alexiadou'01	Picallo'99	Alonso'04	Badia'02
Verbal Class	-	+	+	-	+
Pluralization	+	-	-	+	-
Determiner Type	+	-	+	+	-
Preposition+Agent	-	-	+	-	+
Internal Argument	+	-	+	-	-
External Arguments	+	-	-	-	-
Verbal Predicates	+	-	+	-	+

'as far as'). One of the three denotation values—*event, result, underspecified*—is assigned to the whole lexicalized construction only in the case of nominal lexicalizations. It is important to recognize such lexicalized cases because if the nominalization takes part in a lexicalized construction other than the nominal, it does not receive a denotation (a semantic distinction that is only associated with nouns).

The 3,077 occurrences were classified into 1,121 senses considering that different denotations associated with a lemma are different senses. Henceforth, we refer to them as nominalization senses. Among these 1,121 senses, 807 were annotated as *result* (72%), 113 as *event* (10%), 131 as *underspecified* (12%), and 70 as non-nominal *lexicalized* noun (6%). It is not surprising that *result* nominalizations are the most frequent because *events* tend to be realized mostly by verbal clauses and nominalizations are more frequently used for the *result* (non-dynamic) meaning, more typical of nouns.

The fact that AnCora-ES is annotated with different linguistic information levels allowed for the evaluation of the seven morphosyntactic and semantic criteria selected. Next, we briefly present each criterion, how they were applied, and the results obtained.[10]

4.1.1 Verbal Class. One of the most commonly used criterion to determine the denotation is the verbal class from which the nominalization is derived (Picallo 1999; Alexiadou 2001; Badia 2002). It is claimed that unergative and stative verbs give rise to *result* nominalizations, and unaccusative verbs usually result in ambiguous, or what we call *underspecified*, nominalizations. Regarding transitive verbs, they give rise to either *event*, *result*, or *underspecified* nominalizations. To analyze this criterion, we set out from the semantic verbal classification proposed in AnCora-Verb. In this verbal lexicon, each predicate is associated with one or more semantic classes depending on the four basic types of events (Vendler 1967) (accomplishments which correspond to transitive verbs; achievements corresponding to unaccusative verbs; states corresponding to stative verbs; and activities corresponding to unergative verbs) and on the diathesis alternations in which the verb can occur. We therefore looked up the verbal classes from which the 817 nominalizations are derived in AnCora-Verb. This allowed us to determine whether the claims about the relation between the nominalization denotation type and the corresponding verbal classes are valid.

In the sample analyzed, most of the nominal senses were classified as *results* (72%), thus, it should not surprise us that all the verbal classes have a wide percentage of *result* nominalizations. The most significant result, however, is that stative and unergative verbs lead nearly exclusively to *result* nominalizations in Spanish, 97% and 100%, respectively. Regarding transitive and unaccusative verbs, they lead to *event*, *result*, or *underspecified* nominalizations. It is also interesting to remark that *event* nominalizations derive mostly from transitive verbs (15%, in contrast to the 1% derived from achievement verbs) and *underspecified* nominalizations derive from unaccusative verbs (28%, in contrast to the 11% and 3% derived from transitive and state verbs, respectively), confirming the hypothesis stated by Picallo (1999), Alexiadou (2001), and Badia (2002).

4.1.2 Pluralization. According to Grimshaw (1990) and Alonso (2004), one of the features that clearly identifies *result* nominalizations is their pluralization capacity because it is

10 In the following criteria we do not consider lexicalized senses because the criteria do not apply to these complex lexical units.

more usual to quantify objects than *events*. It is important to point out, however, that it is also possible to make an *event* reading of a plural nominalization. For instance, in Example (11) *bombardeos* ('shelling') refers to multiple actions of bombing, therefore, it is open to an *event* reading.

(11) Los **bombardeos**<event> de Sarajevo por parte del ejército Bosnio.
'The **shelling**<event> of Sarajevo by the Bosnian army.'

This criterion was measured taking into account whether the 817 nominalizations appeared in the plural in some of their occurrences in the sample analyzed. The results obtained (98% of the nominalizations in the plural were classified as *result* and the remaining 2% as *underspecified*) confirmed plurality as one of the features able to detect *result* nominalizations. In contrast, the singular feature is not informative enough to discard any of the nominal denotations (69% of the nominalizations were classified as *results*, 15% as *events*, and 16% as *underspecified* type). Therefore, in the sample analyzed all *event* nominalizations and most of the *underspecified* nominalizations appeared only in the singular.

4.1.3 Determiner Types. Authors such as Grimshaw (1990), Picallo (1999), and Alonso (2004) claim that *event* nominalizations usually appear with definite articles whereas *result* nominalizations may be specified by all types of determiners. For instance, demonstrative determiners can only specify *result* nominalizations because this type of determiner is used to refer to an entity in a frame of reference. In order to evaluate this criterion we took into account the types of determiners combining with the nominalization and also whether the noun appeared without a determiner.

Because most of the nominal senses were classified as *results* (72%), it should not surprise us that all types of determiners have a wide percentage of *result* nominalizations. A striking result observed in Table 3, however, is that indefinite articles (99%), demonstratives (100%), and quantifiers (100%) nearly always appear with *result* senses. In contrast, the definite article, the possessive, and the empty position can occur in all nominalization classes. Seventy-two percent of definite articles appear with *result*, 13% with *event*, and 15% with *underspecified* nominalizations; 82% of possessive determiners appear with *result*, 10% with *events*, and 8% with *underspecified* nominalizations; and 88% of nominalizations without determiner are classified as *result*, 5% as *events*, and 7% as *underspecified* nominalizations. The data therefore partially confirm the hypotheses from the literature: *Result* nominalizations appear with a wider range of determiners. Although *event* nominalizations are not always specified by a definite article, they can also appear with possessive determiners or without any determiner.

4.1.4 Preposition Introducing the Agent. In Spanish nominalizations derived from transitive verbs are considered to be *results* if the *agent* complement is introduced by the preposition *de* ('of') and *events* if the preposition used is *por* ('by') (Picallo 1999).[11] We took into consideration *agent* complements introduced by prepositions appearing in the sample analyzed. As shown in Table 3, Prepositional Phrases (PPs) interpreted as *agents* in the NPs analyzed are introduced by the following prepositions: *de* ('of'), *entre* ('between'), *por* ('by'), and *por parte de* ('by'). We observed that the distribution of the four PPs is complementary between *event* and *result* denotations: The *agent* nominal

11 A similar claim is stated by Badia (2002) for Catalan.

Table 3
Distribution of the denotation types according to the criteria evaluated.

Criteria	Values	Result (%)	Event (%)	Underspecified (%)
Verbal Class	Accomplishments	74	15	11
	Achievements	71	1	28
	States	97	–	3
	Activities	100	–	–
Pluralization	Plural	98	–	2
	Singular	69	15	16
Determiners	Definite	72	13	15
	Indefinite	99	–	1
	Demonstrative	100	–	–
	Possessive	82	10	8
	Quantifier	100	–	–
	No Determiner	88	5	7
Preposition-*Agent*	de *'of'*	98	–	2
	entre *'between'*	100	–	–
	por *'by'*	–	100	–
	por parte de *'by'*	–	100	–
Internal argument	Possessive	41	38	21
	PPs	53	25	22
	Relative Pronoun	71	29	–
	Relational Adjectives	97	–	3
External argument	por *'by'* PPs	–	100	–
	Relational Adjectives	100	–	–
	Possessive	95	–	5
Predicates	Attributive	75	6	18
	Eventive	44	41	15

complement introduced by *de* or *entre* occurs with *result* nominalizations (98% and 100%, respectively) and the *agent* nominal complement introduced by *por* or *por parte de* occurs with *event* nominalizations (100% both), corroborating the hypothesis put forward by Picallo (1999).

4.1.5 Internal Argument. The internal argument criterion proposed by Grimshaw (1990) and Picallo (1999) states that only *event* deverbal nominalizations require the presence of an internal argument because they are more similar to verbs, whereas in *result*-nominalized NPs the internal argument is not needed. Badia (2002) argues that the realization of this argument is not always compulsory to obtain an *event* interpretation of the nominalization. To analyze this criterion, we observed those nominalized NPs in which the internal argument was explicit and the type of argument that realized it. As a result, we observed that the majority of *event* nominalizations are complemented by an internal argument (98%). This is also the case for *underspecified* nominalizations in a fairly high percentage (78%). The percentage decreases considerably in *result* nominalizations (34%), however. Therefore, the data seem to confirm Picallo's and Grimshaw's hypothesis. Table 3 shows that there are four constituents that realize an internal argument: possessive determiners, PPs, genitive relative pronouns, and relational adjectives. The

first two constituents appear in the three nominal denotation types: 41% of possessive determiners appear with *result* nominalizations, 38% with *events* nominalizations, and 21% with *underspecified* nominalizations; and 53% of PPs complement *result*, 25% *events*, and 22% *underspecified* nominalizations. Relative pronouns only occur with *event* (29%) and *result* (71%) denotation types, and, finally relational adjectives occur nearly exclusively in *result* nominalizations (97%). This last fact constitutes an identification feature for *result* nominalizations, as Picallo states (see next criterion).

Table 3 shows the results obtained for each criterion.

4.1.6 External Arguments vs. *Possessors.* Grimshaw (1990) states that PPs introduced by the preposition 'by' (by-PPs), relational adjectives, and possessive determiners in English would be interpreted as external arguments (subjects) in the case of *event* nominalized NPs. These constituents, however, would be interpreted as possessors (that is, as non-argumental) in *result* nominalized NPs. Other authors, like Picallo (1999), nevertheless, claim that possessive determiners may be argumental in *result* and *event* nominalizations in Spanish. Regarding relational adjectives, Picallo argues that these constituents can only be arguments of *result* nominalizations. As seen, there is no consensus among authors; therefore, it seemed to us to be an interesting criterion to contrast. We observed whether these constituents (by-PPs, relational adjectives, and possessive determiners) were interpreted as external arguments in the nominalized NP sample. If this was so, we also analyzed whether the fact of being external arguments conditioned the denotation of the nominalization.[12]

The results obtained are very clear: PPs introduced by *por* ('by') with an *agent* interpretation only occur in NPs headed by *event* nominalizations. Relational adjectives are interpreted as external arguments only in NPs headed by *result* nominalizations. Possessive determiners with an *agent* interpretation are mostly (95%) constituents of NPs headed by *result* nominalizations though they can also be constituents (5%) of NPs headed by *underspecified* nominalizations. Therefore, for Spanish, Grimshaw's hypothesis is only partially corroborated because only by-PPs guarantee the *event* reading. Regarding relational adjectives, Picallo's thesis is confirmed, because this type of constituent mostly appears as an argument of *result* nominalizations. Moreover, we observed a preference for possessive determiners to be external arguments of *result* nominalizations, which is not stated in any of the theoretical proposals.

4.1.7 Verbal Predicates. The type of verbal predicate that can be combined with nominalizations may be an indicator for determining the denotation (Grimshaw 1990; Picallo 1999; Badia 2002). *Result* nominalizations tend to combine with attributive predicates, whereas *event* nominalizations tend to be subjects of predicates such as *tener lugar* ('to take place') or *ocurrir* ('to happen') because these predicates tend to select *event* type subjects. In order to examine this criterion, we analyzed the types of predicates combined with the 817 nominalization types. We observed whether the predicates belong to the *event*-denoting class (*tener lugar*, 'to take place'; *ocurrir*, 'to happen'; *comenzar*, 'to begin'; *acabar*, 'to finish'; *durar*, 'to last'; *llevar a cabo*, 'to carry out') or if they were attributive predicates (*ser*, 'to be'; *estar*, 'to be'; *parecer*, 'to seem'). Table 3 illustrates that

12 The way we decided whether these constituents were external arguments consisted of paraphrasing the nominalized NPs into clause structures in order to see if they were semantically equivalent to verbal subjects.

attributive predicates tend to choose *result* nominalizations (75%) as subjects whereas eventive predicates do not show a clear preference for any type of nominal: Forty-four percent of them combine with *result*, 41% with *events*, and 15% with *underspecified* nominalizations. These results partially confirmed what is stated by these three authors: *result* nominals combine preferentially with attributive predicates.

From the corpus-based study, we conclude that the semantic distinction between *event* and *result* nominalizations is not always as clear as is stated in the literature. The criteria proposed in the literature are well suited to constructed examples but when they are applied to naturally occurring data they do not work so well: Some of them cannot be applied and sometimes we found contradictory criteria in the same example. That said, it is important to point out that these criteria are not irrefutable proofs for making an *event* or a *result* reading, but rather indicators that can help us to strengthen our semantic intuition. In fact, we propose the third denotation type *underspecified* for those cases in which semantic intuition is insufficient and the criteria for reinforcing one of the two main denotation types are not clear.

Regarding the criteria established in the literature, the main conclusion drawn is that not all the criteria analyzed seem to hold for Spanish. Among the evaluated criteria, those that appear to be most helpful for distinguishing between *event* and *result* nominalizations are: 1) the semantic class of the verb from which the noun is derived; 2) the pluralization capacity; 3) the 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 (see Section 5.3).

It is interesting to note that the number of criteria found that reinforce *result* readings is significantly higher than the number of criteria found that strengthen *event* readings. In every criterion we find features that support the identification of *result* nominalizations but not *event* nominalizations. To support *result* nominalizations the following features were found: nominalizations deriving from unergative and stative verbs; nominalizations appearing in the plural; nominalizations specified by an indeterminate article, a demonstrative, or quantifier determiner; nominalizations with an *agent* complement introduced by *de* ('of') or *entre* ('between'); the nonrealization of the internal argument; and nominalizations having relational adjectives as arguments and the attributive predicate combined with them. In order to underpin *event* nominalizations, however, the only unambiguous criterion found was when the preposition introducing a PP *agent* complement is *por* ('by') or *por parte de* ('by').[13] If we take into account that the *agent* complement is mostly optional in an NP configuration, it is very difficult to find a criterion within the NP context to support *event* nominalizations.

We believe that there are more features to support *result* nominalizations because they are closer to non-derived nouns and, like them, admit a wide variety of configurations: plural inflection, different types of determiners, the possibility of appearing without complements, and so forth. In contrast, *event* nominalizations (since they are not typical nouns because they denote an action), like verbs, do not admit this variety of configurations: They rarely appear without complements, admit fewer types of determiners, and appear in the plural less frequently. Most of the configurations they admit are also admitted by *result* nominalizations; this explains why there are more criteria to support *result* than *event* nominalizations.

13 Literally, 'on the part of.'

In fact, the remaining criteria—the fact of deriving from transitive or unaccusative verbs; the nominalization being in the singular; the co-occurrence with a definite article, a possessive determiner, or without any determiner; the presence of the internal argument; and the combination with typically eventive predicates—do not support any specific denotation. As a result, there are several cases where it is very difficult to assign a denotation, especially when the context is not clear enough, and therefore, we need to apply the *underspecified* tag.

In the next section, we present other indicators found in the empirical study that provide us with clues for the differentiation between *event* and *result* nominalizations. These indicators are data-driven and we can only guarantee that they work for Spanish.

4.2 Finding New Clues to Support *Event* and *Result* Denotations

The analysis of 3,077 nominalization occurrences, focusing on the semantic distinction between those denoting an *event*, *result*, or *underspecified* type, has allowed us to find new clues that strengthen these readings, especially the *event* reading.

One of the clearest clues for detecting the *event* nominalizations is the possibility of paraphrasing the NP with a clausal structure, as we saw in Section 1, Examples (3)–(6). Another valuable clue is to check whether the nominalization admits an *agent* complement introduced by *por* ('by') or *por parte de* ('by'). We use this criterion because it is the most informative one to support *event* nominalizations but it is also a very optional complement and is scarcely represented in the corpus. The annotators could use these two tests to decide the denotation type. Therefore, they had two extra criteria that the data did not provide.

Furthermore, we found other indicators that can help to select one denotation type, the so-called **selectors**. We identified two types of selectors:

1. External selectors: Prepositions like *durante* ('during'), nouns like *proceso* ('process'), adjectives like *resultante* ('resulting'), verbs like *empezar* ('begin'), and adverbs *en vía de* ('on the way to'), which are elements that point to a specific denotation from outside the nominalized NP. For instance, in Example (12) the preposition *durante* ('during') gives a clue to interpret *presentación* ('presentation') as an *event*.
2. Internal selectors: Prefixes within the nominalization that indicate a specific denotation type; for instance, a noun with the prefix *re-* with a reiterative meaning such as *reubicación* ('relocation') in Example (13). This is due to the fact that the reiterative meaning only applies to bases that denote actions.

(12) Durante [la **presentación**<event> del libro], él abogó por la formación de los investigadores en innovación tecnológica.
'During [the **presentation**<event> of the book], he advocated the training of researchers in technological innovation.'

(13) Hoy [la **reubicación**<event> del ex ministro] no resulta fácil.
'Today, [the **relocation**<event> of the ex minister] does not seem easy.'

These new clues allow us to support our semantic classification independently from the literature criteria under evaluation. The only inconvenience of these tests and the selectors is that they cannot be implemented as features in the ADN-classifier.

5. Knowledge Resources

This section presents the linguistic resources used in building the final version of the ADN-Classifier (R3). We briefly describe the AnCora-ES corpus and the AnCora-Verb lexicon, and we focus in more detail on the description of the AnCora-Nom lexicon from which we obtain most of the features for the building of the ADN-Classifier.

5.1 AnCora-ES Corpus

AnCora-ES is a 500,000 word (henceforth, 500kw) Spanish corpus[14] consisting of newspaper texts annotated at different linguistic levels: morphology (part of speech and lemmas), syntax (constituents and functions), semantics (verbal argument structure, thematic roles, semantic verb classes, named entities, and WordNet nominal senses), and pragmatics (coreference).[15] The corpus contains 10,678 fully parsed sentences. As we explained in Section 3, nominalization occurrences (23,431) were automatically annotated with denotation types using an intermediate model of classification (see step 5 in Figure 1). This automatic annotation was then manually validated by three graduate students in linguistics. These annotators were selected from a group of five, because they achieved an observed agreement of over 80%, corresponding to a kappa of 65% in an inter-annotator agreement test, whereas the average observed agreement was 75% corresponding to a 60% kappa. For the purpose of annotation, the three annotators took into account the semantic definition we provided, the criteria presented in Section 4.1, and the semantic tests described in Section 4.2. The inter-annotator agreement was carried out to ensure the consistency and quality of the AnCora-ES manual annotation.[16]

Therefore, the AnCora-ES corpus enriched with denotation type annotation is used for learning the different models of the ADN-Classifier-R3. From this resource we obtained two kinds of features:

(a) The corpus versions of the features from the lexicon (see Section 5.3): the type of determiner used in Section 4; the number (plural or singular) in which the nominalization occurrences appear; and the constituent type of the complements.

(b) The contextual features such as the tense and the semantic class of the verb that dominates the nominalization in the sentence; the syntactic function of the NP headed by a nominalization; and whether the noun appears in a named entity.

We use the Tgrep2[17] tool for the feature extraction from the corpus; this allows us to efficiently inspect the syntactic trees in a Treebank format.[18]

14 A similar version exists for Catalan, AnCora-CA.

15 AnCora-ES is the largest multilayer annotated corpus of Spanish freely available at: `http://clic.ub.edu/corpus/ancora`.

16 For more details on the manual validation and the inter-annotator agreement test, see Peris, Taulé, and Rodríguez (2010).

17 `http://tedlab.mit.edu/ dr/TGrep2/`. Tgrep2 is an improvement of Tgrep. Both tools are tree-based counterparts of the widely used string searching Unix Grep tool.

18 In the following link the set of tgrep rules as well as some implemented examples are available: `http://clic.ub.edu/corpus/en/documentation`.

5.2 AnCora-Verb

AnCora-Verb-ES is a verbal lexicon that contains 2,830 Spanish verbs.[19] In this lexicon, each predicate is related to one or more semantic classes, depending on its senses, basically differentiated according to the four event classes—accomplishments, achievements, states, and activities (Vendler 1967; Dowty 1979)—and on the diatheses alternations in which a verb can occur (Vázquez, Fernández, and Martí 2000). The semantic class of the verb base of the nominalization is used as a feature in the ADN-Classifier.

5.3 AnCora-Nom

This section presents AnCora-Nom,[20] a Spanish lexicon of deverbal nominalizations that has been iteratively used and improved as a result of the experiments reported here. At present, it contains 1,655 lexical entries corresponding to 3,094 senses and 3,204 frames. These lexical entries represent the lemmas corresponding to the 23,431 deverbal nominalization occurrences appearing in the annotated AnCora-ES corpus. For each of these lemmas we created a lexical entry using the information annotated in the corpus.[21] The features of each lexical entry are organized in three levels: lexical entry, sense, and frame level. The lexical entry attributes are not extracted from the corpus but added in order to document the lexical entry. Sense and frame attributes, in contrast, were extracted from the AnCora-ES corpus. Each lexical entry is organized in different senses, which were established taking into account the denotation type, the sense of the base verb, and whether or not the nominalization is part of a lexicalized construction. In turn, each sense can also contain one or more nominal frames, depending on the verbal frame from which the nominalization is derived. Next, we detail the attributes specified in the three levels described above. Figure 2 shows the full information associated with the lexical entry *aceptación* (‘acceptance’).

5.3.1 Lexical Entry Level Attributes. These are as follows:

(a) **Lemma**. In Figure 2, the value for this attribute is the noun *aceptación* (lemma=“aceptación”).

(b) The attribute **language** (“lng”) codifies the language represented in the lexical entry. AnCora resources work with Spanish and Catalan, so the values of this attribute are “es” for Spanish (lng=“es”) and “ca” for Catalan (lng=“ca”). At present, AnCora-Nom only deals with Spanish nominalizations.

(c) The attribute **origin** indicates the type of word from which the nouns are derived. In Figure 2, the value for this attribute is “deverbal”, meaning that this lexical entry concerns a noun derived from a verb. At present, AnCora-Nom only contains deverbal nouns but in the future it will include other types of nominalizations such as deadjectivals.

(d) The attribute **type** refers to the word class, “noun” in Figure 2.

19 A similar version exists for Catalan, AnCora-Verb-CA.
20 We describe here AnCora-Nom-R3, the final version of the lexicon.
21 For a detailed explanation of the automated extraction process see Peris and Taulé (2011).

```
<?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" postype="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>

...     </examples>

</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" postype="possessive" type="determiner"/>

</argument>

<specifiers>

<constituent frequency="2" postype="article" type="determiner"/>

</specifiers>

<examples>

<example file="CESS-CAST-A/11714_20000314.tbf.xml" nodepath="7.4.1.1.1.1.3.2.1.2.0" sentencenodepath="7"> En el
marco de esta estrategia marcada por la prudencia , el PP esperará a los movimientos que haga el consejero de Economía ,
Artur_Mas , 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>
```

Figure 2
Aceptación (‘acceptance’) lexical entry in AnCora-Nom.

5.3.2 Sense Level Attributes. These include:

(e) The attribute **cousin** marks whether the nominalization is morphologically derived from a verb (cousin="no", in Figure 2) or is a cousin noun (cousin="yes"). Cousin nouns (Meyers, Reeves, and Macleod 2004) are nouns that give rise to verbs (e.g., *relación* ['relation'] >*relacionar* ['to relate']), or nouns semantically related to verbs (e.g., *escarnio* ['mockery'] is related to *mofarse* ['to make fun']).

(f) The **denotation** attribute indicates the semantic interpretation of the deverbal noun. The possible values are: "event," "result," and "underspecified." In Figure 2, there are two senses, the first one being *result* (denotation="result") and the second one *event* (denotation="event").

(g) Each sense contains an **identifier** ("id") to indicate the sense number.

(h) The **lexicalized** attribute indicates whether or not the nominalization is part of a lexicalized construction (idiomatic expression) (Figure 2: lexicalized="no"). In the first case, two additional attributes are added: (i) the **alternative-lemma**, specifying the whole lexicalized construction of which the nominalization is part (for instance, alternative-lemma="golpe de estado," ['coup d'etat']), and (ii) **lexicalization-type**, to distinguish between the six types of lexicalizations: "nominal," "verbal," "adjectival," "adverbial," "prepositional," or "conjunctive" (see Section 4). We should bear in mind that one of the three above-mentioned denotation values is assigned to the whole lexicalized construction only in the case of nominal lexicalizations. For instance, the lexicalized construction *golpe de estado* is a nominal lexicalization (lexicalization-type= "nominal"), and therefore, it has a denotation value (denotation="result").

(i) The attribute **originlemma** specifies the verb lemma from which the noun is derived. In Figure 2, the value for this attribute is "aceptar" in both senses (originlemma="aceptar").

(j) Because verbs can have different senses, the attribute **originlink** indicates the concrete verbal sense of the base verb. In Figure 2, the **originlink** attribute takes the same value in both senses: "verb.aceptar.1" (oringinlink="verb.aceptar.1").

(k) Because nouns in the AnCora corpus are annotated with WordNet synsets,[22] we incorporate this information in the attribute **wordnetsynset**. In Figure 2, the first sense of *aceptación* corresponds to two synsets (wordnetsynset="16:00117820+16:10039397"), whereas the second only corresponds to one (wordnetsynset="16:00117820"). It should be noted that senses in AnCora-Nom are coarser grained than in WordNet: a sense can group together more than one WordNet synset.

5.3.3 Frame Level Attributes. These are detailed as follows:

(l) The attribute **type** indicates the verbal frame from which the nominalization is derived. In AnCora-Verb, each verbal sense can be realized in more than one frame: default, passive, anticausative, locative, and so forth. In the nominal entries, we mark the corresponding verbal frames, which are the possible values for this attribute. In most cases, its value is "default" as in Figure 2 (type="default"). This feature is needed to look for the corresponding verbal semantic class in AnCora-Verb.

22 We used WordNet 1.6 for Spanish and WordNet offsets for identifying synsets.

(m) **Argument (Structure)**. In this complex attribute, the different arguments (**argument**) and the corresponding thematic roles (**thematicrole**) are specified. To represent the arguments we follow the same annotation scheme used in AnCora-Verb. For instance, in Figure 2, the *event* sense has one argument ("arg1") with a patient thematic role ("pat"). This argument is realized twice (frequency="2") by a prepositional phrase (constituent type ="sp") introduced by the preposition *de* ('of') (preposition="de") and once by a possessive determiner (type="determiner," postype="possessive").

(n) The attribute **referencemodifier** represents the nominal complements that are not arguments but which modify the reference of the nominalization. Frequency is also taken into account. Strictly speaking, this attribute does not fit perfectly at the frame level. We were interested in representing this information, however, and the most suitable level was the frame level because it allows for a seamless comparison of argumental and nonargumental nominal complements.

(o) The type of determiner has proved to be a useful criterion for distinguishing between *result* and *event* readings, so we include this information in the attribute **specifier**.[23] The possible values are: "article," "indefinite," "demonstrative," "exclamative," "numeral," "interrogative," "possessive," "ordinal," and "void" when there is no determiner. In this attribute, we also take into account the frequency with which the determiners are realized. In Figure 2, the *event* sense is specified twice (constituent frequency="2") by an article determiner (type="determiner," postype="article").

(p) The attribute **appearsinplural** indicates whether or not an occurrence of a nominalization in a particular frame appears in the plural. It is a boolean attribute. In Figure 2, neither of the senses appear in the plural, thus, the value is "no."

(q) Finally, each lexical entry also contains all the **examples** from which the information has been extracted, specifying the **corpus file**, the **node path**, and the **sentence** in which each is located.

6. ADN-Classifier

As stated previously, our goals for building the ADN-Classifier were twofold: On the one hand, to have at our disposal a tool to help us to quantitatively evaluate the validity of our claims regarding deverbal nominalizations as discussed in Section 4; and, on the other hand, to provide a classification tool able to take advantage of all the available information in a specific scenario in the automatic classification of a deverbal noun. The aim of the task is to classify a deverbal nominalization candidate in an *event*, *result*, or *underspecified* denotation type, as well as to identify whether the nominalization takes part in a *lexicalized* construction (idiomatic expression). Therefore, we model the task as a four-way classification problem. In order to achieve these goals, some functional requirements on the software to be built were necessary. Regarding the first goal, we required that a tool be able to:

1. Use all the properties discussed in Section 4 as features for classification.

23 The name of the attribute refers to the syntactic position that determiners occupy in the NP; the determiners specify the nominalization.

2. Tune the features: binarization, grouping the possible values, generalization, combination of features, and computation of derived features.
3. Perform feature selection.
4. Facilitate the human interpretation of the model used by the classifier.
5. Analyze the accuracy of the individual features.
6. Use either senses or lemmas corresponding to deverbal nominalization candidates as units for classifying.

In order to achieve the first aim, the first version of the ADN-Classifier (R1) (Peris, Taulé, and Rodríguez 2009) was developed. This is the basis for the building of the intermediate and final versions of the ADN-Classifier. The final version is presented in detail next. Obviously, the second goal imposes heavier constraints on the design of the tool (the ADN-Classifier-R3). As is usual in other lexical classification tasks, such as Part Of Speech (POS) tagging or WSD, a word taken as an isolated unit is ambiguous but can be disambiguated, or at least partially disambiguated, if enough context is taken into account. An additional constraint is that the nominalization candidate has to be tagged as a noun. For our classification task at least four processes are carried out: 1) tokenization; 2) segmentation at sentence level; 3) POS tagging; and 4) localization of a nominalization candidate by means of a set of regular expressions looking for verbal nominalization endings.[24]

In this setting, a case for classifying consists of a nominalization candidate using the POS-tagged sentence where it occurs as context, although this context is not always sufficient for disambiguation. Additional processes could be carried out on the nominalization candidate and the sentence (WSD, chunking, full parsing, SRL, linking of the nominalization candidate with the origin verb, etc.). Each of these processes increases the number of possible features used for classifying but, because they are not error free, they could involve a decrement in the global accuracy of the preprocess step. Therefore, a careful examination of the processes, their accuracy, and the improvement of classification accuracy is needed. For instance, performing WSD on the nominalization candidate could allow for the use of sense-based features and, thus, a clear improvement in classification accuracy. The inconvenience is that the state-of-the-art accuracy rate of WSD is not very promising. In recent SemEval challenges, the accuracy rate in All-Words tasks is between 60% and 70% for a baseline of 51.4% using the first sense in WordNet, and 89% in Lexical-Sample tasks for a baseline of 78% (Chklovski and Mihalcea 2002; Decadt et al. 2004; Pradhan et al. 2007).[25] These figures depend on the sense inventory used for disambiguation: The All-Words task uses fine-grained senses (WordNet synsets) and the Lexical Sample task uses more coarse-grained sense inventory (Ontonotes senses).

Therefore, we approach the problem of classification taking into account different feature sets which come from different knowledge resources, and we examine and evaluate the task performance when a decreasing number of knowledge resources are used. Depending on the available knowledge resources and natural language

24 We used 10 suffixes such as *-ción* (see Section 4).

25 All these figures are for English. To have some idea of the relative difficulty of the task for Spanish we have measured this baseline in Ancora-ES resulting in a value of 42%, that is, an 18% drop with respect to English.

Table 4
Description of the scenarios used for evaluation.

Scenario	Knowledge Resources	Features level	NL Pre-Process
1	AnCora-Nom+AnCora-Verb	lemma	POS
2	AnCora-Nom+AnCora-Verb	sense	POS+WSD
3	AnCora-Nom+AnCora-Verb	lemma	POS+Parsing
4	AnCora-Nom+AnCora-Verb	sense	POS+WSD+Parsing
5	AnCora-Nom	lemma	POS
6	AnCora-Nom	sense	POS+WSD
7	AnCora-Nom	lemma	POS+Parsing
8	AnCora-Nom	sense	POS+WSD+Parsing
9	–	lemma	POS+Parsing
10	–	lemma	POS+Parsing+SRL

(NL) processors, we designed the classification task in different scenarios, which are presented in Table 4. The columns include the knowledge resources used in each scenario (column 2), whether the features used are extracted at sense or lemma level (column 3), and the NL processors that are necessary in each case.

Scenario 1 in Table 4 presents the case in which the nominal lexicon (AnCora-Nom in our case) is available and the nominalization candidate is an entry in this lexicon. The sentence where the nominalization candidate occurs is POS-tagged and no other NL processes are carried out.[26] In this case, we apply the ADN-Classifier with a model learned using only features coming from the lexicon at the lemma level with $Acc_{lemma;lex}$ accuracy. Scenario 2 is the same as Scenario 1 but adds a WSD process to the nominalization candidate with Acc_{WSD} accuracy. In this case, we apply the ADN-Classifier with a model learned using only lexicon features at a sense level achieving $Acc_{sense;lex}$ accuracy. Obviously, applying this model is only useful if $Acc_{lemma;lex} - Acc_{sense;lex}$ outperforms the expected WSD error ($1 - Acc_{WSD}$). Scenario 3 is the same as Scenario 1 but adds constituent parsing with Acc_{parser} accuracy. In this case, we apply the ADN-Classifier with a model learned using lexicon and corpus features[27] at lemma level with $Acc_{lemma;lex+corpus}$ accuracy. Again, this model is only useful if the $Acc_{lemma;lex+corpus} - Acc_{lemma;lex}$ outperforms the expected parsing error ($1 - Acc_{parser}$). Scenario 4 consists of a combination of Scenarios 2 and 3. Scenarios 5, 6, 7, and 8 reproduce Scenarios 1, 2, 3, and 4, respectively, without using the features extracted from the AnCora-Verb lexicon, so obtaining the origin verb of the candidate is not necessary. In Scenario 9, the nominal lexicon is not available or the nominalization candidate is a noun that does not occur in the nominal lexicon, and only the features extracted from the parsed tree at lemma level are used. Finally, Scenario 10 is the same as Scenario 9 but adds an SRL process in order to obtain argument structure information. Taking into account these two sets of requirements the final version of the ADN-Classifier (R3) has been built.

We used ML techniques to build the ADN-Classifier. Specifically, we used the J48.Part rule-based classifier, the rule version of the decision-tree classifier C4.5.Rules (Quinlan 1993) as implemented in the Weka toolbox (Witten and Frank 2005). We chose a rule-based classifier because it provides a natural representation of classification rules, thus allowing for the inspection of the model without diminishing accuracy and it

26 POS-tagging implies previous tokenization and sentence segmentation steps.
27 The parse tree obtained can be inspected in the same way as AnCora-ES.

allows us to perform a ranking of the individualized rules and a definition of a thresholding mechanism for performing a precision oriented classification. Because most of the features are binary and the others are discrete with small range values, using more complex rule-based classifiers such as Cohen's Ripper supposes no real improvement over our choice.[28]

The ADN-Classifier therefore consists of the J48.Part classifier within the Weka toolbox, the appropriate learned model (from the set described in Section 7.2 and listed in Table 5), and the list of features to be used. During the exploitation phase the input to the system consists of a table. Each row in the table corresponds to a case for classifying and each column to the values of the corresponding feature. The result of the process is a column vector containing the result of classifying each instance.

7. Experiments and Evaluation

In this section we present and evaluate the experiments carried out with the ADN-Classifier. First we present the settings of these experiments, then we focus on the experiments themselves, and finally we evaluate the results.

7.1 Setting

In order to validate the performance of the ADN-Classifier, a sequence of experiments was conducted. Concretely, two sets of experiments were carried out: we experimented with different models of the classifier structured in five dimensions (see Section 7.2) and we applied the appropriate models in the different scenarios set out in Table 4. We use a tenfold cross-validation method[29] for the evaluation of these two sets of experiments. In order to evaluate the features selected and to carry out the classification task in each scenario we used the models learned as described in Section 7.2. As noted earlier, using the ADN-Classifier for classify involves using the J48.Part classifier within the Weka toolbox and the appropriate learned model.

7.2 Experiments

The experiments carried out with the ADN-Classifier-R3 are presented here. Firstly, we describe those experiments related to the different models of the classifier and secondly, we focus on how some of these models are applied in different scenarios. We apply the ADN-Classifier in different modes that correspond to the following five dimensions.

28 J48.Part learns first a decision tree and then builds the rules traversing all the branches of the tree. Ripper, instead, learns the rules one by one (increasing the learning cost). This can result in a more accurate and smaller rule set just in the case of splitting numerical attributes; that is not our case.

29 In n-fold cross-validation, the original sample is randomly partitioned into n subsamples. Of the n subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $n - 1$ subsamples are used as training data. The cross-validation process is then repeated n times (the folds), with each of the n subsamples used exactly once as the validation data. The n results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. n is commonly set to 10 (McLachlan, Do, and Ambroise 2004).

Table 5
Experiments and evaluation of the models. Legend of the model name: 1st letter (S = sense-based, L = lemma-based); 2nd letter (S = sense, L = lemma, E = corpus example); 3rd letter (L = lexicon features, C = corpus features, A = all features); 4th letter (R = reduced vocabulary, F = full vocabulary) and 5th letter (R = reduced corpus, F = full corpus).

		Model	Inst.	Att.	Rules	Baseline (%)	Accu. (%)	Δerror (%)	RedΔerror (%)
Sense-based models	Senses	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
	Examples	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
		SECFF	23,431	197	211	85.45	**84.93**	−0.52	−3.60
		SEARR	1,840	1,133	76	85.32	91.57	6.25	42.59
		SEARF	9,278	1,133	196	87.03	96.38	9.35	72.15
		SEAFR	3,994	1,867	146	83.92	91.72	7.80	48.52
		SEAFF	23,431	1,867	498	85.45	**95.46**	10.01	68.83
Lemma-based models	Lemmas	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
	Examples	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
		LECRR	1,840	197	35	85.32	84.02	−1.30	−8.88
		LECRF	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
		LECFF	23,431	197	211	85.45	**84.86**	−0.58	−4.04
		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

– **Application level**. We distinguish between sense-based and lemma-based models. Sense-based models use the information from the AnCora-Nom-R3 lexicon at sense level, that is, the features for learning (and classification) are associated with the specific senses. In contrast, in lemma-based models, when extracting features from the lexicon, we use as features for learning and classification those attributes whose values are shared by all senses of the same lemma. Therefore, at this second level of application the features are not so informative but, at the same time, we reduce our dependence on the lexicon, which was a step that had to be taken to move towards a more realistic scenario.

– **Unit of learning and classification** (i.e., the instance to be classified). These sense- or lemma-based models are in turn distinguished depending on whether the unit of learning and classification comes from the lexicon (sense or lemma) or from the

AnCora-ES corpus (examples), that is, if they correspond to types or tokens. In the first case all the features come from the lexicon, whereas in the latter contextual features, extracted from the corpus, can also be used. Consequently, in sense-based models the units used are senses (from the lexicon) or examples (from the corpus), and in lemma-based models they are lemmas (from the lexicon) or examples (from the corpus). It has to be taken into account that depending on this dimension, the number of instances for learning varies: There are more senses than lemmas in the lexicon and there are more nominalization occurrences (examples) in the corpus than nominalization senses or lemmas in the lexicon.

– **Features involved**. The features used for both learning and classifying are obtained from the lexicon (lexical features) or from the corpus (contextual features). The different models are distinguished by using only lexical features, only contextual features, or the combination of both types of features.

– **Vocabulary size**. The data sets taken into account correspond to a reduced set of 817 nominalization lemmas obtained from the 100kw subset of the corpus used for the first version of the ADN-Classifier (R1) or to the full set of 1,655 nominalization lemmas occurring in the whole AnCora-ES (500kw). Depending on this dimension, the number of instances for learning also vary.

– **Corpus size**. Two corpus sets are used in the different models: a reduced subset of 100kw used for the first version of the ADN-Classifier (R1) or the full 500kw corresponding to the whole corpus.[30] Depending on this dimension, the number of instances for learning also vary.

In order to identify the models as presented in Table 5, we use five letters as notation, each of which identifies one of these five dimensions. The first letter corresponds to the application level: If the model is sense-based an S will be used for the identification and an L in the case of lemma-based models. The second letter refers to the unit of classification and is L for lemmas, S for senses, and E for examples. In the third position the reference to the origin of the features involved in the model is found: L (from the lexicon), C (from the corpus), A (from both resources, all features). In fourth place, we refer to the vocabulary size: R (reduced) stands for the reduced set of 817 nominalization lemmas and F (full) for the full set of 1,655 nominalization lemmas. In last place, we also designate the corpus size by an R (reduced set of 100kw) or an F (full set of 500kw). Therefore a lemma-based model that uses examples as units of classification, uses all the features, the whole vocabulary, and the whole corpus is identified as the LEAFF model. In total we experimented with 32 different models.[31]

For the different scenarios described in Section 6, we applied the appropriate models so as not to use the noninformed features for each one.

7.3 Evaluation

The classifier performance of the different models was evaluated by a tenfold cross-validation method. Next, we focus on the results of the 32 models resulting from the five dimensions described in Section 7.2. Table 5 presents the overall results: the models

30 For obtaining the learning curve of some of our models intermediate sizes have been used.
31 Not all the combinations of values of the dimensions are allowed.

used (column 1), the number of instances used for learning (column 2), the number of attributes used, and the number of rules built by the classifier (columns 3 and 4), and finally, the baseline, the accuracy, the decrease error over the baseline (Δ-error), and the relative error-reduction ratio (Red-Δ-error) obtained by each model (columns 5, 6, 7, and 8). The rows correspond to the different models presented. Recall that the names of the models are assigned according to the five dimensions presented in Section 7.2. It should be borne in mind here that in column 2, the number of instances for learning depends on the type of unit used for learning and classification (senses in sense-based models, lemmas in lemma-based models, and examples) and on the vocabulary and corpus size. The interaction between these three dimensions also explains why the figures for the baseline change for each model. The baseline is a majority baseline which assigns all the instances to the *result* class. In general, when the unit used is from the lexicon, the lemma baseline increases relative to the sense baseline. This is because in lemma-based models we group the senses that share all the features under a lemma; because different senses do not normally share all the features, in the end, only monosemic lemmas are in fact taken into account. This fact, therefore, shows that there are more *result* type monosemic lemmas than *event* and *underspecified* monosemic lemmas. Furthermore, it is worth noting that when the unit of learning and classification used are the examples from the AnCora-ES corpus and not the senses from the lexicon, the baseline also increases. Therefore, it seems that proportionally *result* nominalizations are more highly represented in the corpus than *event* and *underspecified* nominalizations. Regarding the number of features used for learning, the type of feature involved and the vocabulary size (when features from the lexicon are used) are the two relevant dimensions. Finally, it should be said that the accuracy and the other two correlated measures are obtained by evaluating the performance of the different models by tenfold cross-validation.

As can be seen in Table 5, the sense-based models (the first 16 rows) outperform the corresponding lemma-based models (the last 16 rows). This is explained by the fact that there are features in the lexicon coded at the sense level that cannot be recovered at the lemma level because in lemma-based models we only use as features for the classification those attributes whose values are shared by all senses of the same lemma, and this does not commonly happen. At the sense level, the best results are achieved when the features used in the classification come exclusively from the lexicon, with the unit of classification being senses from the lexicon (the first block of four rows) or examples from the corpus (the second block of four rows). The contextual features (those coming from the corpus) can only be applied to models using examples from the corpus as the unit of classification. These features harm accuracy: When they are used alone (the third block of rows) they yield accuracy values that are below the baseline and when they are used in combination with features obtained from the lexicon (the fourth block of rows) the accuracy decreases in relation to the models that use only the lexicon as the source of the features (the second block of four rows). This shows that there is crucial information in the lexicon that is not possible to recover from the corpus. Furthermore, it should be mentioned that there is a generalized improvement across sense-based models correlated to the vocabulary and especially corpus size: The bigger the set of vocabulary and corpus, the better the result. This fact is also present in lemma-based models.

The sense-based models represent the upper bounds for our task. In a realistic scenario, however, given the state-of-the-art results in WSD, we would not have access to sense labels, so we are much more interested in the performance of lemma-based models. The best results are achieved when features from the lexicon and from the

corpus are combined (the last block of rows), showing that the sum of both types of features gives rise to positive results, which are not achieved by lexical features or contextual features on their own. When the features used in the classification come exclusively from the lexicon, with the unit of classification being lemmas from the lexicon (the fifth block of four rows) or the examples from the corpus (the sixth block of four rows), the results are negative (below the baseline) except when the vocabulary and corpus size are both the full sets (1.54% and 0.16% improvement, respectively). In these cases, the information from the lexicon is not as accurate as in sense-based models. The contextual features alone do not achieve positive results, not even with the full vocabulary and the full corpus. Therefore, the combination of features is needed in a realistic scenario in order to achieve good performance of the classifier. In these cases, only when the reduced vocabulary and the reduced corpus are used are the results slightly negative. From now on, we will focus on the last model (LEAFF) because even if it has a lower accuracy than the corresponding sense-model, we expect it to exhibit a more robust behavior when tackling unseen data.

An important point for the classifier to learn a model is whether or not the sample size is large enough for accurate learning. We performed a learning curve analysis of the LEAFF model for different sample sizes (from 1,000 examples to the whole set of 23,431 examples). The results are depicted in Figure 3. We have also plotted the confidence intervals at 95%. The results seem to imply that for sizes over 5,000 examples the accuracy tends to stabilize; we are, therefore, highly confident of our results. As expected, the confidence intervals consistently diminish as the corpus grows.

7.4 Precision Oriented Classifier: Threshold

All the experiments reported so far are based on a full coverage setting. Coverage is 100% in all cases and, therefore, accuracy and precision have the same score.

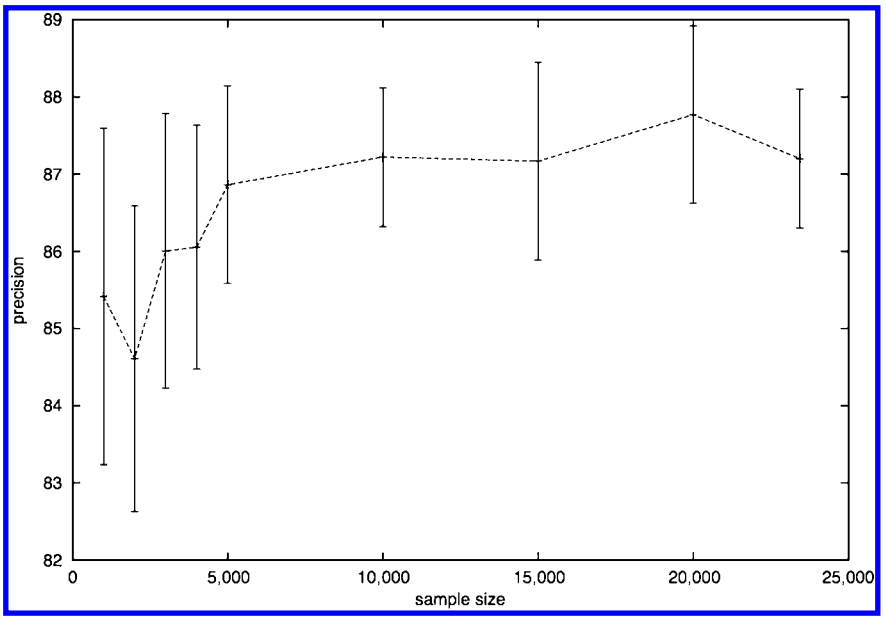

Figure 3
Learning curve for the LEAFF model.

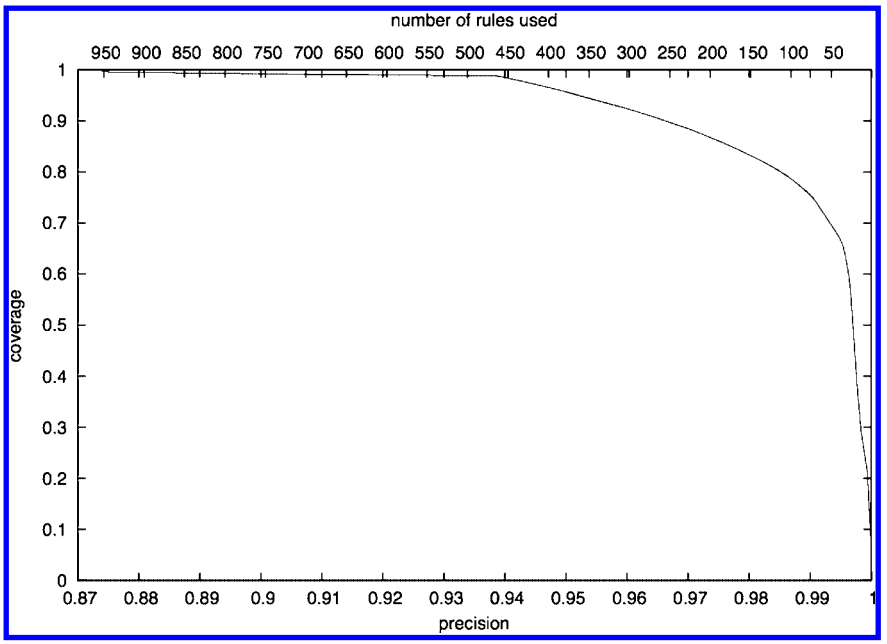

Figure 4
Precision/coverage plot for different rules thresholding of the SEAFR model.

Additionally, we performed a precision oriented experiment based on a classifier aiming to achieve a high precision at the cost of a drop in coverage. In order to do this, we scored each of the rules in the rule set from the LEAFF model of the ADN-Classifier individually (not taking into account the order of such rules). We sorted the rules in the rule set by their individual scores, provided by the Weka toolbox, and built a classifier based on a thresholding mechanism: Only the rules over the threshold were applied. This resulted in higher precision at the cost of lower coverage. The LEAFF model, as presented in Table 5, consists of 981 rules obtaining an overall accuracy of 87.20%. The results of the application of the n most accurate rules, for n from 981 to 1, are depicted in Figure 4. Note that removing the 500 least accurate rules has a small effect on the coverage and the overall precision has risen to 94%. An additional reduction of 200 rules results in an increase of the overall precision to 96.5% at a cost of a drop in the coverage to 90%. Using only the 100 most accurate rules obtains a precision of 98.5% for a still useful coverage of 80%.

7.5 Evaluation of Scenarios

The results of the experiments applying the scenarios described in Section 6 (see Table 4) are presented in Table 6. The table shows the results of the ten scenarios set out in rows, and in columns we provide the scenario identification (column 1); the model applied out of the 32 generated, following the notation in Section 7.2 (column 2); the number of features in the original model (column 3); the number of features in the model adapted for that scenario after removing noninformed features, that is, the features used in the original model that do not fit in the description of a concrete scenario (column 4); and the accuracies of the original and final model (columns 5 and 6, respectively). In each scenario we applied the best model of the 32 we generated taking into account

Table 6
Experiment and evaluation of scenarios.

Scenario	Model	Initial Att.	Final Att.	Initial Acc. (%)	Final Acc. (%)
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	LECFF	197	197	84.86	84.86
10	LEAFF	1,755	1,556	87.20	87.08

the features that each model uses and that fit the best in each scenario according to the hypothesized available linguistic processors. When there is no concrete model to project how the ADN-Classifier would perform in a concrete scenario, we selected the model that fits approximately in that scenario and removed the noninformed features. For instance, Scenario 10 describes the case where the nominal lexicon is not available or the nominalization candidate is a noun that does not occur in the nominal lexicon, and the features used are extracted from the parsed tree at lemma level and from the SRL process in order to obtain argument structure information. Because we do not have a model that perfectly fits in that scenario, we select the LEAFF (lemma based model using examples from the corpus as the unit of classification and obtaining the features from both lexicon and corpus, with full vocabulary and full corpus sets), and we removed all the features from the lexicon except the ones related to the argument structure, simulating an SRL process.[32]

These results show that the difference between lemma-based and sense-based models shown in Table 5 is also present here. There is a decrease in accuracy in all the cases in which features are removed, although this decrease is not statistically significant. This could be due to the large number of features available for rule learning and the possibility of using alternate features when some of the original ones are removed.

7.6 Error Analysis

The analysis of errors focuses on the lemma-based model using lexicon and corpus information with the full vocabulary and the full corpus (LEAFF). Table 7 presents the confusion matrix of the model. Rows correspond to manually labeled data and columns are predictions from the classifier. The correct predictions are in the diagonal (in **boldface**). The errors are marked in *italics*.

The rate of error is almost equally split between the three main classes: incorrectly classified *event* nominalizations represent 31%, *result* nominalizations 34%, and *underspecified* nominalizations 32%. *Lexicalized* instances,[33] however, only display an error rate of 3%. Among *event* nominalizations incorrectly classified by the ADN-Classifier,

32 In the same way, sense-based models simulate how the ADN-Classifier would perform with an ideal WSD automatic process.

33 By lexicalized nominalizations we refer here to those lexicalized nominalizations where a denotation type is not assigned, that is, non-nominal lexicalizations.

Table 7
Confusion matrix for the LEAFF model.

ADN		Result	Event	Underspecified	Lexicalized	Total
	Result	**18,997**	*575*	*397*	*54*	20,023
Manually	Event	*676*	**933**	*242*	*2*	1,853
Validated	Underspecified	*643*	*309*	**453**	*7*	1,412
	Lexicalized	*90*	*2*	*2*	**49**	143
	Total	20,406	1,819	1,094	112	23,431

73% (676 instances) were classified as *result*; 26% (242 instances) as *underspecified*, and a marginal 3% (2 instances) as lexicalized nominalizations. These errors are attributable to four main causes. Firstly, 27% of the errors are in fact human errors,[34] which means that the ADN-Classifier performed well. Secondly, the annotation guidelines contain criteria that the ADN-Classifier cannot recognize: the paraphrase criterion, the *agent* criterion, and the so-called *selectors* (see Section 4.2). These errors represent 51% of the wrongly classified *events*. Therefore, there were cases (a total of 61) where manual annotators had an extra criterion that the ADN-Classifier could not use. We thought that implementing the selectors as features in the ADN-Classifier would be an excessively ad hoc approach. Thirdly, an error of 21% in *event* classification is explained because there are a number of criteria, implemented as features in the ADN-Classifier, that suffer from data sparseness, and, therefore, the ADN-Classifier cannot learn them as being as relevant as they are. For instance, a very conclusive clue for detecting *event* nominalizations is that they are specified by a possessive determiner that is interpreted as an arg1-patient. And, finally, the cases in which the ADN-Classifier annotated *event* nominalizations as *lexicalized* constructions are explained by the ADN-Classifier confusing them with real lexicalized constructions in which the lemma is shared (an error rate of 1%).

Among *result* nominalizations incorrectly classified by the ADN-Classifier, 56% (575 instances) were classified as *event*, 39% (397 instances) as *underspecified*, and 5% (54 instances) as lexicalized nominalizations. These errors are attributable to the same four causes set out above. The rate of human errors is now 51%, however, meaning that there are *event* and *underspecified* nominalizations that were incorrectly validated. The rate of errors explained by the selectors is now just 10% because there are more selectors for identifying *event* than for detecting *result* nominalizations. And finally, an error rate of 37% is explained by those criteria that are implemented as features of the ADN-Classifier, but which suffer from data sparseness, and, therefore, despite their relevance, cannot be learned by the ADN-Classifier. In the case of *result* nominalizations, there are more criteria of this type: nominalizations deriving from unergative and stative verb classes, relational adjectives as arguments, and temporal arguments realized by a PP introduced by *de* ('of'). And finally, the cases in which the classifier annotated *result* nominalizations as *lexicalized* constructions are explained by the ADN-Classifier confusing them with real lexicalized constructions in which the lemma is shared (an error rate of 2%).

34 When comparing automatic annotation with the manual validation, some cases were considered to be doubtful. We discussed those cases with all the annotators and decided which annotation (automatic or human) was the correct one. Therefore, by human errors we mean those incorrectly classified in the manual validation process.

Among incorrectly classified *underspecified* nominalizations, 32% (309 instances) were classified as *events*, 67% (643 instances) as *results*, and a marginal 1% (7 instances) as *lexicalized* nominalizations. The difficulty in identifying *underspecified* nominalizations is to be expected, given that these are either cases with no clear contextual hints or truly ambiguous examples. In this case, the rate of human error is 45%. Although there are no selectors that identify *underspecified* nominalizations, in some cases an NP containing a nominalization presents contradictory criteria. For instance, an indefinite determiner is a criterion that points to a *result* denotation and the selector *durante* ('during') typically selects an *event* denotation. In these cases, the annotators were instructed to tag them as *underspecified*. The ADN-Classifier could not use the selectors in its classification, however, and most of these cases therefore were annotated as *results*. This type of error represents 19% of the incorrectly classified *underspecified* nominalizations. The *agent* criterion explains an error rate of 20%. If both types of PPs (introduced by the preposition *por* ['by'] or introduced by the preposition *de* ['of']) are valid for the NP the annotators were validating, they tagged the nominalization as *underspecified* type. Again, human annotators had an extra criterion that the ADN-Classifier could not use. The remaining 5% is explained by the failure of the ADN-Classifier to detect a pattern that is typical of *underspecified* nominalizations: those derived from an achievement verb with an arg1-patient. And finally, the cases in which the ADN-Classifier annotated *underspecified* nominalizations as *lexicalized* constructions are explained by the ADN-Classifier confusing them with real lexicalized constructions in which the lemma is shared (an error rate of 1%).

Most incorrectly classified *lexicalized* constructions (96%, 90 instances) were classified as *result* nominalizations. This is probably due to the fact that most nominal lexicalized nominalizations are of the *result* type. Therefore, the key failure of the ADN-Classifier is basically in distinguishing between the different types of lexicalized constructions.

8. Related Work

Although there are several works that contemplate the computational treatment of nominalizations, most of them are basically interested in two issues: 1) automatically classifying semantic relations between nominals (Task 4 of SemEval 2007 [Girju et al. 2009] and Task 8 of SemEval 2010 [Hendrickx et al. 2010]) or in noun compound constructions (Girju et al. [2005] and Task 9 of SemEval 2010 [Butnariu et al. 2010a, 2010b]); and 2) taking advantage of verbal data to interpret, represent, and assign semantic roles to complements of nominalizations (Hull and Gomez 2000; Lapata 2002; Padó, Pennacchiotti, and Sporleder 2008; Gurevich and Waterman 2009). Although most of these works show a certain awareness of the linguistic distinction between *event* and *result* nominalizations, none of them applies this distinction in their systems. The notion of *event* appears in the work of Creswell et al. (2006), in which a classifier that distinguishes between nominal mentions of *events* and *non-events* is presented. Their distinction is not comparable to our *event* and *result* distinction for one main reason, however: they do not focus on nominalizations but on nouns in general, and therefore the difficulty in distinguishing *events* from *non-events* among all types of nouns is less than distinguishing between *event* and *result* nominalizations, which, as has been seen, are highly ambiguous. We state that it is easier because in a wide nominal domain there are many nouns which under any circumstance can be understood as non-dynamic (in fact, nouns tend to denote objects, *non-events*) and if Creswell et al. include as seed for

learning these types of nouns, such as *airport* or *electronics*, it will necessarily increase the accuracy of the automatic distinction between their two denotation types.

As far as we know, the only work closely related to ours is that of Eberle, Faasz, and Ulrich (2011) who are working on German *–ung* nominalizations, in which the assignment of a denotation type is also carried out. In that paper they state that this kind of nominalization can denote an *event*, a *state*, or an *object*. Specifically, they analyze those *–ung* nominalizations derived from verbs of saying embedded in PPs introduced by the preposition *nach* ('to'). According to the authors, these nominalizations can denote either an *event* or a *proposition*, which is a type of *object*. They present a semantic analysis tool (Eberle et al. 2008) which disambiguates this type of nominalization on the basis of nine criteria, which they call indicators. The tool extracts the indicators from the semantic representation that it provides and computes the preferred denotation according to a pre-established weighting schema. They apply this tool to a set of 100 sentences where the relevant material (the nine indicators) is completely familiar to the system and the tool recognizes the preferred reading in 82% of cases.

Because the ADN-Classifier is based on ML techniques and does not restrict the nominalizations to a specific suffix and to those derived from verbs of saying, their work is not directly comparable to ours. We tried to replicate their experiment, however, selecting only those nominalizations created with the suffix *-ción* (the most productive Spanish suffix and the nearest equivalent to *–ung* in German) and which derive from verbs of saying. The subset obtained includes 66 types of nominalizations, compared with the 1,655 in our work. We applied the LEAFF model to the 719 tokens of these 66 nominalization types and we obtained an accuracy of 85.6%. This implies a 3.6 percentage point increase in accuracy with respect to the results of their work, despite the fact that our model is not trained on this specific nominalization class and does not dispose of specially designed indicators. We have to take this result with due caution because we are dealing with two not closely related languages and considering a close but not identical set of nominalizations.

9. Conclusions

This article contributes to the semantic analysis of texts focusing on Spanish deverbal nominalizations. We base our study on theoretical hypotheses that we analyze empirically, and as a result we have developed three new resources: the ADN-Classifier, the first tool that allows for the automatic classification of deverbal nouns as *event* or *result* types; the AnCora-ES corpus enriched with the annotation of deverbal nominalizations according to their semantic denotation, being in fact the only Spanish corpus which incorporates this information; and the AnCora-Nom lexicon, a resource containing 1,655 deverbal nominalizations linked to their occurrences in the AnCora-ES corpus. These resources could be very useful for NLP applications. The work presented in this article also provides an additional insight into the linguistic question underlying it: The characterization of deverbal nominalizations according to their denotation and the identification of the most useful criteria to distinguish between these denotation types. We can classify our contributions in three directions:

1) **The study of the relevant features for the classification of a nominalization as being of *event* or *result* type.** The set of features considered were selected from the linguistics literature, mostly devoted to the English language, and its relevance was established empirically for Spanish. From the corpus-based study, we concluded that not all the criteria posited for English seem to port 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. 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 note that the number of features used to support *result* nominalizations is significantly superior to those used to strengthen *event* nominalizations. In each criterion we find features for supporting *result* nominalizations but not *event* nominalizations. As a result, the ADN-Classifier uses more features for detecting *result* than *event* nominalizations, and therefore achieves a greater degree of accuracy on the former than in the latter. Furthermore, the corpus base study has allowed us to find new clues that support denotation types, especially the *event* reading. The paraphrase and agent criteria, as well as the selectors, have proved very useful to human annotators for distinguishing between an *event* and a *result* reading. These criteria are difficult to implement automatically, however.

2) **Lexical resources derived from this work.** We have enriched the AnCora-ES corpus with the annotation of 23,431 deverbal nominalization occurrences according to their semantic denotation; and we have built AnCora-Nom from scratch, representing the 1,655 nominalization types that correspond to these occurrences.

3) **The ADN-Classifier.** This 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 claims regarding deverbal nominalizations. An error analysis was performed and its conclusions can be used to pursue further lines of improvements.

Further work. Two of the main sources of error found in the performance of the ADN-Classifier 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 Classifier. Among them are path-based syntactic patterns that have been applied with success 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. Most of the considerations regarding the scenarios described in Table 4 are based on a crude global evaluation of complementary NL processors such as a word sense disambiguator; for

example, a specific scenario can be followed when the global accuracy of the NL processor crosses a given threshold. A more precise approach can also be adopted. Consider, for instance, the WSD task instead of a simple classifier providing a global accuracy—a regressor can provide individual scores of accuracy for each case (degree of confidence, margin, probability of correct classification, etc.). This more precise approach can lead to new scenarios incorporating hybrid models.

The last point of future work consists in analyzing to what extent the ADN-Classifier and its models are applicable to other languages. Concretely, because 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.

Acknowledgments

We are grateful to Maria Antònia Martí and Marta Recasens for their helpful advice and to David Bridgewater for the proofreading of English. We would also like to express our gratitude to the three anonymous reviewers for their comments and suggestions to improve this article. This work was partly supported by the projects Araknion (FFI2010-114774-E), Know2 (TIN2009-14715-C04-04), and TEXT-MESS 2.0 (TIN2009-13391-C04-04) from the Spanish Ministry of Science and Innovation, and by a FPU grant (AP2007-01028) from the Spanish Ministry of Education.

References

Abeillé, Anne, Lionel Clément, and Alexandra Kinyon. 2000. Building a treebank for French. *In Proceedings of the Second International Language Resources and Evaluation (LREC'00)*, pages 87–94, Athens.

Alexiadou, Artemis. 2001. *The Functional Structure in Nominals. Nominalizations and Ergativity*. John Benjamins, Amsterdam/Philadelphia, PA.

Alonso, Margarita. 2004. *Las construcciones con verbos de apoyo*. Visor Libros, Madrid.

Androutsopoulos, Ion and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Aparicio, Juan, Mariona Taulé, and M. Antònia Martí. 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.

Badia, Toni. 2002. Els complements nominals. In Joan Solà, editor, *Gramàtica del Català Contemporani*, volume 3. Empúries, Barcelona, pages 1591–1640.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 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.

Balvet, Antonio, Lucie Barque, and Rafael Marín. 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.

Bos, Johan. 2008. Wide-coverage semantic analysis with Boxer. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 277–286, Venice.

Butnariu, Cristina, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010a. 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-2010)*, pages 100–105, Boulder, CO.

Butnariu, Cristina, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010b. 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.

Chklovski, Timothy and Rada Mihalcea. 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.

Civit, Montserrat and Maria Antònia Martí. 2004. Building Cast3LB: A Spanish

Treebank. *Research on Language and Computation*, 2(4):549–574.
Clark, Herbert H. 1975. Bridging. In *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '75, pages 169–174, Stroudsburg, PA.
Copestake, Ann. 2007. Semantic composition with (Robust) Minimal Recursion Semantics. In *Proceedings of the Workshop on Deep Linguistic Processing*, DeepLP '07, pages 73–80, Prague.
Creswell, Cassandre, Matthew J. Beal, John Chen, Thomas L. Cornell, Lars Nilsson, and Rohini K. Srihari. 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.
Decadt, Bart, Véronique Hoste, Walter Daelemans, and Antal Van Den Bosch. 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.
Dowty, David. 1979. *Word Meaning and Montague Grammar*. Reidel, Dordrecht.
Eberle, Kurt, Gertrud Faasz, and Ulrich Heid. 2009. Corpus-based identification and disambiguation of reading indicators in German nominalizations. In *Online Proceedings of the 5th Corpus Linguistics Conference*. Available at `ucrel.lancs.ac.uk/publications/cl2009/`.
Eberle, Kurt, Gertrud Faasz, and Heid Ulrich. 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, Kurt, Ulrich Heid, Manuel Kountz, and Kerstin Eckart. 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*, pages 145–158, Konvens.
Erk, Katrin and Sebastian Padó. 2006. Shalmaneser: A flexible toolbox for semantic role assignment. In *Proceedings of the Fifth International Language Resources and Evaluation (LREC'06)*, pages 527–532, Genoa.
Fellbaum, Christiane. 1998. *An Electronic Lexical Database*. The MIT Press, Cambridge, MA.
Gerber, Matthew and Joyce Y. Chai. 2010. Beyond NomBank: A study of implicit argumentation for nominal predicates. In *Proceedings of the ACL Conference 2010*, pages 1583–1592, Uppsala.
Gildea, Daniel and Martha Palmer. 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.
Girju, Roxana, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer, Speech and Language*, 19(4):479–496.
Girju, Roxana, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter D. Turney, and Deniz Yuret. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105–121.
Grimshaw, Jane. 1990. *Argument Structure*. The MIT Press, Cambridge, MA.
Gurevich, Olga, Richard Crouch, Tracy Holloway King, and Valeria De Paiva. 2006. Deverbal nouns in knowledge representation. In *Proceedings of Florida Artificial Intelligence Research Society Conference*, pages 670–675, Florida.
Gurevich, Olga and Scott Waterman. 2009. Mapping verbal argument preferences to deverbals. In *Proceedings of the 2009 IEEE International Conference on Semantic Computing*, pages 17–24, Washington, DC.
Hendrickx, Iris, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 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.
Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of Human Language Technologies–North American Chapter of the Association of Computational Linguistics (HLT-NAACL' 06)*, pages 57–60, New York, NY.
Hull, Richard D. and Fernando Gomez. 2000. Semantic interpretation of deverbal nominalizations. *Natural Language Engineering*, 6(2):139–161.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1027–1032, Genova.

Lapata, Maria. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press Inc., New York.

Liu, Chang and Hwee Tou Ng. 2007. Learning predictive structures for semantic role labeling of NomBank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 208–215, Prague.

Madnani, Nitin and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Márquez, Lluís, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

McLachlan, G. J., K. A. Do, and C. Ambroise. 2004. *Analyzing Microarray Gene Expression Data*. Wiley, Hoboken, NJ.

Mechura, Michal. 2008. *Selectional Preferences, Corpora and Ontologies*. Ph.D. thesis, Trinity College, University of Dublin.

Meyers, Adam, Ruth Reeves, and Catherine Macleod. 2004. NP-external arguments: A study of argument sharing in English. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE'04)*, pages 96–103, Stroudsburg, PA.

Mooney, Raymond 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, Berlin.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Nunes, Mary L. 1993. Argument linking in English derived nominals. In Robert D. Van Valin, editor, *Advances in Role Reference Grammar*. John Benjamins, Amsterdam/Philadelphia, pages 375–432.

Padó, Sebastian, Marco Pennacchiotti, and Caroline Sporleder. 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.

Palmer, Martha, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling. Synthesis on Human Languages Technologies.* Morgan & Claypool Publishers, Toronto.

Palmer, Martha, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):76–105.

Peris, Aina and Mariona Taulé. 2011. AnCora-Nom: A Spanish lexicon of deverbal nominalizations. *Procesamiento del Lenguaje Natural*, 46:11–19.

Peris, Aina, Mariona Taulé, Gemma Boleda, and Horacio Rodríguez. 2010. ADN-Classifier: Automatically assigning denotation types to nominalizations. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1422–1428, Valleta.

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

Peris, Aina, Mariona Taulé, and Horacio Rodríguez. 2010. 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.

Picallo, Carme. 1999. La estructura del Sintagma Nominal: las nominalizaciones y otros sustantivos con complementos argumentales. In Ignacio Bosque and Violeta Demonte, editors, *Gramática Descriptiva de la Lengua Española*, volume 1. Espasa Calpe, Madrid, pages 363–393.

Pradhan, Sameer, Honglin Sun, Wayne Ward, James H. Martin, and Dan Jurafsky. 2004. Parsing arguments of nominalizations in English and Chinese. In *Proceedings of Human Language Technologies–North American Chapter of the Association of Computational Linguistics (HLT-NAACL' 04*, pages 208–215, Boston, MA.

Pradhan, Sameer S., Edward Loper, Dmitriy Dligach, and Martha Palmer. 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.

Pustejovsky, James. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Recasens, Marta, M. Antònia Martí, and Mariona Taulé. 2007. Text as scene: Discourse deixis and bridging relations. *Revista de la Asociación Española para el Procesamiento del Lenguaje Natural, 39*, pages 205–212.

Recasens, Marta and Marta Vila. 2010. On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. FrameNet II: Extended theory and practice. Technical report, International Computer Science Institute, Berkeley, CA. Available at `framenet.icsi.berkeley.edu/book/book.pdf`.

Ruppenhofer, Josef, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 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.

Santiago, Ramón and Enrique Bustos. 1999. La derivación nominal. In Ignacio Bosque and Violeta Demonte, editors, *Gramática Descriptiva de la Lengua Española*, volume 3. Espasa Calpe, Madrid, pages 4505–4594.

Spencer, Andrew and Marina Zaretskaya. 1999. The Essex database of Russian verbs and their nominalizations. Technical report, University of Essex, Colchester.

Surdeanu, Mihai, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 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.

Taulé, Mariona, M. Antónia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 96 –101, Marrakech.

Vázquez, Glòria, Ana Fernández, and Maria Antònia Martí. 2000. *Clasificación Verbal. Alternancias de Diátesis*. Quaderns de Sintagma, 3, Edicions de la Universitat de Lleida, Lerida, Spain.

Vendler, Zeno. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA.

Zubizarreta, Maria Luisa. 1987. *Levels of Representation in the Lexicon and in the Syntax*. Foris, Dordrect.