Iarg-AnCora: Spanish corpus annotated with implicit arguments

Mariona Taulé · Aina Peris · Horacio Rodríguez*

Centre de Llenguatge i Computació (CLiC), University of Barcelona.

Gran Via, 585, 08007 Barcelona, Spain

Tel: 0034934035671 Fax: 0034933189822

mtaule@ub.edu, aina.peris@ub.edu

*TALP Research Center, Technical University of Catalonia.

Jordi Girona Salgado 1-3, 08034 Barcelona, Spain

Tel: 0034934017024 Fax: 003434934017014

horacio@lsi.upc.edu

Abstract

This article presents the Spanish Iarg-AnCora corpus (400 k-words, 13,883 sentences) annotated with the implicit arguments of deverbal nominalizations (18,397 occurrences). We describe the methodology used to create it, focusing on the annotation scheme and criteria adopted. The corpus was manually annotated and an interannotator agreement test was conducted (81% observed agreement) in order to ensure the reliability of the final resource. The annotation of implicit arguments results in an important gain in argument and thematic role coverage (128% on average). It is the first corpus annotated with implicit arguments for the Spanish language with a wide coverage that is freely available. This corpus can subsequently be used by machine learning-based semantic role labeling systems, and for the linguistic analysis of implicit arguments grounded on real data. Semantic analyzers are essential components of current language technology applications, which need to obtain a deeper understanding of the text in order to make inferences at the highest level to obtain qualitative improvements in the results.

Keywords

Implicit Argument, Deverbal nominalizations, Argument Structure, Thematic Roles, Semantic Corpus annotation, Linguistic Resource

1 Introduction

Iarg-AnCora[1] is the result of enriching the AnCora-Es corpus (Taulé et al. 2008; Recasens and Martí 2010) with the addition of a new layer of semantic annotation: the implicit arguments of deverbal nouns (i.e. arguments that are not expressed syntactically in the local context of these predicates, whose semantic interpretation depends on the linguistic and extralinguistic context). For instance, in sentence (1), the arguments of the Spanish deverbal noun *operación* ('operation' or 'surgical operation') are not explicitly realized in the sentence, but one of them, *el paciente* ('the patient'), the entity to be operated on, can be recovered from the previous sentence.

[1] Iarg-AnCora is freely available at: http://clic.ub.edu/corpus/en/ancora-descarregues

AnCora-Es had only been annotated with the explicit arguments of deverbal nouns (23,439 nominal tokens) (Peris and Taulé 2012) and verbal predicates (56,590 verbal tokens) (Taulé, Martí and Recasens 2008). Therefore, it had only been tagged with the arguments appearing within the noun phrase (NP) in the case of nouns, or within the sentence in the case of verbs. Iarg-AnCora integrates the annotation of both explicit and implicit arguments. We focused on deverbal nominalizations because the arguments tend to be expressed implicitly (60% of the cases in the corpus) rather than realized locally in the NP (38% of the cases in the corpus) in this type of predicate. In the case of verbs, most of the arguments occur explicitly within the verbal phrase (VP) (1.32 explicit arguments vs. 0.19 implicit arguments per verb on average[2]) (See Table 11 for details). In example (1), the patient argument of the verb *operar* ('perform surgery'), that is, *al paciente* ('on a patient'), is explicitly realized within the VP, while the patient argument of the noun *operación* is not explicitly realized.

(1) [No han llegado los productos necesarios para [*operar* [al paciente]]$_{VP}$]$_{sentence}$. [Por lo tanto, [la *operación*]$_{NP}$ se ha cancelado]$_{sentence}$.
'[The products needed [*to perform surgery* [on a patient]]$_{VP}$ have not arrived in time]$_{sentence}$. [Therefore, [the operation]$_{NP}$ has been cancelled]$_{sentence}$ '.

We can, therefore, postulate that the degree of optionality of the explicit arguments of deverbal nominalizations is higher than for verbs (see section 5), and that to fully understand the meaning of deverbal nouns it is necessary to take into account both explicit and implicit arguments. Since verbs contain more explicit arguments, disposing of the implicit ones is not so critical for obtaining their correct meaning. Due to the limited resources available for annotating the corpus, a costly task, we considered that the annotation of both verbs and nominalizations could not be undertaken and that annotating only the deverbal nominalizations would result in a more valuable resource. In any case, our intention is to enrich Iarg-AnCora with the implicit arguments of verbal predicates in a future work.

This resource is an important contribution to the semantic analysis of texts due to the scarcity of corpora annotated with implicit arguments, most of which are created for English (Ruppenhofer et al. 2010; Gerber and Chai 2012; Moor, Roth and Frank 2013) and have restricted coverage (1,250-3,000 occurrences of nominal and verbal predicates), or need to be extended with artificial examples in order to tackle the problem of sparseness in Machine Learning (ML) tasks, as shown in Table 1 and discussed in section 3. Iarg-AnCora is the only Spanish corpus that is annotated with the argument structure of deverbal nominalizations, and the only corpus freely available with such a wide coverage (18,397 deverbal noun occurrences). Iarg-AnCora will be a valuable resource to train and test Semantic Role Labeling (SRL) systems and Semantic Parsers using ML techniques, as well as to infer linguistic knowledge about the way the implicit arguments of deverbal nominalizations occur in real data. In fact, a previous version of this corpus was used to train and test the LIARc classifier (Labeling Implicit ARguments in Spanish deverbal nominalizations) (Peris et al. 2013). The corpus could also prove interesting in order to study discourse coherence and its modeling (Roth and Frank 2013).

Another important strength of Iarg-AnCora is that it integrates different levels of semantic and discourse annotation: argument structure and thematic roles (for both verbs and deverbal nouns), named entities, Spanish WordNet nominal and verbal senses, and coreference. A corpus with all this semantic-discursive information integrated is undoubtedly an interesting and useful resource for semantic analyzers whose aim is a deeper understanding of the text in order to make inferences on the highest level and thereby obtain qualitative improvements in the results. It is a language resource that can be used for many Natural Language Processing tasks and applications that need to go beyond shallow parsing, such as Question Answering, Information Extraction and Machine Translation.

(2) [[La *operación* [financiera]$_{AP}$ [de este proyecto]$_{PP}$]$_{NP}$ no cuenta con la aprobación del Banco de España.] $_{sentence}$ [Por lo tanto, [la *operación*]$_{NP}$ no se pudo realizar]$_{sentence}$.
'[[The [financial]$_{AP}$ *transaction* [for this project]$_{PP}$]$_{NP}$ has not been approved by the Spanish National Bank.]$_{sentence}$ [Therefore, [the *transaction*] $_{NP}$ could not be carried out]$_{sentence}$.

[2] Since implicit arguments are not annotated in AnCora-Es, the percentage of realization cannot be computed. The corresponding figure (0.19 implicit arguments per verb) has been estimated from the corpus assuming that for a given predicate the number of arguments (explicit or not) is the same, on average, when realized as a verb or as deverbal nominalization.

In a Machine Translation system, the deverbal noun *operación* in the second sentences of examples (1) and (2) would be ambiguous if we did not take into account the information contained in the previous sentence. However, the recovery of the implicit arguments -*el paciente* 'the patient' in (1) and *financiera* 'financial' in (2)- allows us to translate *operación* in (1) as 'operation' and *operación* in (2) as 'transaction'.

This resource may also be of particular interest for SRL systems, especially for those dealing with the explicit arguments of both nominal and verbal predicates and also for those taking into account the implicit arguments of deverbal nouns. It could also be of interest for Coreference Resolution (CR) systems because the linking of implicit arguments to their antecedent can be considered as a special case of coreference (Silberer and Frank 2012). In fact, the Implicit Semantic Role Labeling (ISRL) is a task that combines techniques of SRL and CR.

The article is organized as follows. We first introduce the notion of implicit argument (section 2), and related work (section 3). Then we describe the methodology carried out to create Iarg-AnCora (section 4), presenting the annotation scheme and the criteria adopted, the linguistic resources used, the results of the interannotator agreement test conducted, and the annotation interface used. Next, we present some statistics on the annotation (section 5), and we briefly describe the use of Iarg-AnCora for learning and testing the LIARc classifier (section 6). Finally, conclusions are drawn (section 7).

2 Implicit arguments

An implicit argument, or a null instantiation (NI) in terms of the FrameNet framework (Fillmore 1986, Fillmore and Baker 2001), is an argument syntactically unrealized in the local context of the predicates (verbs, adjectives or nouns) whose semantic interpretation depends on the linguistic or extralinguistic context. The implicit argument can be either a core argument (which represents an essential participant in the action/event evoked by the predicate) or an adjunct (optional) argument. Adjunct arguments are optional by definition and have a limited impact on the semantic interpretation of the predicate. Moreover, predicates usually place few and loose discriminative constraints on their adjunct arguments. Research dealing with implicit arguments therefore focuses on core arguments. In FrameNet, NIs are further classified as definite null instantiations (DNI) -anaphorically bound within the discourse- or indefinite null instantiations (INI) -existentially bound within the discourse. However, Moor, Roth and Frank (2013) prefer to distinguish between resolvable and non-resolvable NIs within the discourse, rather than classifying them as DNIs and INIs. Gerber and Chai (2010, 2012) do not take into account the distinction between INIs and DNIs.

In this work, we only detect and classify the implicit core arguments of the deverbal nominalizations whose semantic interpretation depends on the linguistic context, which is, in our case, the whole document. If the semantic interpretation depends on the extralinguistic context, it cannot be recovered from the surrounding discourse (3) and (4). In contrast, if the sentence interpretation depends on the linguistic context, implicit arguments can be recovered and linked to an entity (5) and (6). In the case of nominalizations, implicit arguments can be recovered from the sentence containing the nominalization or from the previous or following sentences (6). In this sense, Iarg-AnCora was only annotated with the resolvable implicit arguments of deverbal nominalizations, that is, resolvable NIs as defined by Moor, Roth and Frank, or the DNIs proposed by FrameNet.

In the following examples we provide insights into the use of explicit and implicit arguments both in the cases in which the predicate is realized as a verb and as a deverbal nominalization. Although a detailed account of our annotation scheme is presented in section 4.2, a short excerpt of our notation is included here in order to allow for the interpretation of the examples below. We use the symbol [Ø] to indicate the presence of an implicit argument. We annotate the implicit arguments with the *iargn="r:entity$_x$"* tag, with the letter *i* standing for an implicit argument (explicit arguments are annotated without a prefix), *n* for an argument position and *r* for a thematic role. When the implicit argument can be recovered it is linked to an underlined[3] entity identified with a number *entity$_x$* (x=identifies the entity number).

Examples (3) and (5) are cases in which the involved predicates are verbs. Although only deverbal

[3] For sake of clarity we underline the discourse entities acting as antecedents of the implicit arguments.

nominalizations are annotated with implicit arguments in Iarg-AnCora, we include these examples for the sake of completeness. In examples (4), (6), and (7) the involved predicates are deverbal nouns, the focus of our work.

(3) No hay duda de que [*se cometieron* [errores]<arg1="pat"> [Ø]<**iarg0="agt"**>]][4].
'There is no doubt that [[mistakes]<arg1="pat"> *were made* [Ø]<**iarg0="agt"**>]'.

(4) Todavía no sabemos nada sobre [la *oferta* [de la petrolera]<arg0**="agt"**> [Ø]<**iarg1**="pat">]NP.
'We do not know anything about [the oil company's<arg0**="agt"** *offer* [Ø]<**iarg1**="pat">]NP yet'.

(5) [El vuelo a Estambul<**entity="entity1"**> se retrasó] y [*llegamos* [Ø]<**iarg3:"dest:entity1"**> [a las tres de la tarde]<argM="tmp">].
'[The flight to Istanbul<**entity="entity1"**> was delayed] and [*we arrived* [Ø]<**iarg3:"dest:entity1"**> [at three in the afternoon]<argM="tmp">]'.

(6) En el seno de la directiva barcelonista<**entity="entity1"**>, se ha remplazado el silencio por [las *apariciones* [Ø]<**iarg1:"tem:entity1"**> [públicas]<arg2:"loc">]NP para no desanimar aún más al barcelonismo<**entity="entity2"**>. Sin embargo, existe [un manifiesto *desencanto* [Ø]<**iarg1:"tem:entity2"**> [por el futuro del Barcelona]<arg0="cau">]NP.
'Within the Barcelona board<**entity="entity1"**>, silence has been replaced [by [public]<arg2:"loc"> *appearances* [Ø]<**iarg1:"tem:entity1"**>] in order not to further discourage the fans<**entity="entity2"**>. However, there is [serious disenchantment [Ø]<**iarg1:"tem:entity2"**> [with the future of Barcelona]<arg0="cau">]NP'.

In example (3), the agent implicit argument (iarg0="agt") of the verbal predicate *se cometieron* ('were made') cannot be recovered (or resolved) within the linguistic context (it is an INI in FrameNet terminology), therefore, this argument ([Ø]<iarg0="agt">) is not linked to a discourse entity. The agent argument is indefinite, existentially bound within the discourse; we do not know who exactly committed the mistakes, because it is probably irrelevant in communicative terms. This is a general grammatical characteristic of passive constructions, in which the patient argument is nearly always explicitly realized (arg1="pat"), though not the agent argument. Another example of a non-resolvable implicit argument (in terms of Moor, Roth and Frank) of a deverbal noun is (4), where the implicit patient argument ([Ø]<iarg1="pat">), in this case the specific offer made, cannot be recovered from the linguistic context and cannot, therefore, be linked to an entity in the discourse such as in (3). In Iarg-AnCora, these non-resolvable arguments are not annotated.

However, in example (5) the implicit destination argument ([Ø]<iarg3="dest:entity1">) of the verbal predicate *llegamos* ('we arrived') can be recovered from the linguistic context (it is a DNI in FrameNet terminology). Concretely, it can be linked to *Estambul*<*entity=entity1*> ('Istanbul') from the previous sentence. Example (6) shows the deverbal noun *apariciones* ('appearances') with the explicit locative argument (arg2="loc") (*públicas*, 'public') realized inside the NP, whereas the theme argument ([Ø]<iarg1="tem:entity1">) (*la directiva barcelonista*<*entity=entity1*>, 'the Barcelona board') is implicitly understood and recovered from the same sentence but outside the NP. Instead, in the same example (6), the implicit theme argument ([Ø]<iarg1="tem:entity2">) (*el barcelonismo*<*entity=entity2*>, 'the fans') of the deverbal noun *desencanto* ('disenchantment') can be recovered from the previous sentence. In the case of sentence (6), the identification of the implicit arguments implies a gain in the semantic role coverage of two deverbal nouns (*apariciones* 'appearances' and *desencanto* 'disenchantment'), and therefore a gain for the semantic interpretation of these sentences and for the understanding of the text.

It should also be pointed out that in Iarg-AnCora corpus we treat those arguments that are syntactically unrealized in the local context of these predicates, i.e. the arguments realized outside the NP, as implicit arguments of deverbal nouns. But, unlike Meyers (2007), we also treat those that do not depend directly on the nominal predicate, even though they appear in the NP, as implicit arguments. In other words, the implicit arguments can occur within the scope of a nominal predicate without being directly dominated by it. For instance, constituents inside a subordinate clause of the deverbal noun can be implicit arguments (7).

[4] In section 4.1, the annotation scheme is presented in more detail.

(7) [El *daño* [Ø]<iarg0="cau"> [Ø]<iarg1="tem:entity1"> [causado a su industria aeronáutica<entity="entity1">]$_S$[5]]$_{NP}$.
'[The *damage*[Ø]<iarg0="cau"> [Ø]<iarg1="tem:entity1"> [caused to its aeronautics industry<entity="entity1">]$_S$]$_{NP}$'.

In example (7), the implicit theme argument ([Ø]<iarg1="tem:entity1">) of the deverbal noun *daño* ('damage') is the *industria aeronáutica*<entity="entity1"> ('aeronautics industry'), which is a constituent inside the subordinate clause (*S*). It is not an explicit argument because the deverbal noun does not directly dominate it. In the case of the implicit causative argument ([Ø]<iarg0="cau">), more linguistic context is needed in order to recover the referent of the argument.

3 Related work

There exist different corpora that are semantically annotated with the argument structure and thematic roles of deverbal nominalizations: NomBank[6] (Meyers, Reeves and Macleod 2004; Meyers 2007) and FrameNet[7] (Ruppenhofer et al. 2006) reference resources for English; AnCora-Es[8] (Peris and Taulé 2012) for Spanish; NOMAGE[9] (Balvet et al. 2011) for French, and the on-going Copenhagen Dependency Treebank (CDT)[10] project (Müller 2011), which aims to semantically annotate a set of parallel treebanks for Danish, English, German, Italian and Spanish. However, all of them are focused on the annotation of explicit arguments, that is, those arguments realized inside the NP which includes the nominalization. Corpora annotated with the implicit arguments of deverbal nouns are very scarce, restricted to English, and have limited coverage.

As far as we know, there are two English corpora annotated with implicit arguments specifically created to train and test SRL systems, which are presented in Ruppenhofer et al. (2010) and in Gerber and Chai (2010, 2012). The former corresponds to the training and test corpus developed for SemEval-2010 task 10, *Linking events and their participants in discourse*[11] (Ruppenhofer et al. 2010, 2012). The corpus consists of literary texts extracted from two of Arthur Conan Doyle's fictional works, annotated following the FrameNet-style (Erk and Padó 2004)[12]. The number of nominal and verbal occurrences tagged is 3,073 (corresponding to 769 different frame types) in a total of 963 sentences (17,072 tokens). Therefore, in this corpus each nominal and verbal predicate has a very small number of occurrences (an average of 4 instances per predicate), and data is consequently rather sparse. In fact, this scarcity of data is one of the reasons put forward by Gerber and Chai (2010, 2012) for the creation of a new dataset for developing and evaluating their SRL system. They tagged a subset of the standard training, development and testing sections of the Penn TreeBank (Marcus et al. 1993) following the PropBank (Palmer, Kingsbury and Gildea 2005) and NomBank (Meyers 2007) annotation scheme. In order to avoid the problem of sparseness, they tagged a large number of occurrences (1,247 in total) of only 10 different nominal predicates. The predicates chosen correspond to the ten[13] most frequent unambiguous deverbal nouns in the corpus. Following a similar approach, Moor, Roth and Frank (2013)[14] tagged the implicit arguments of 1,992 occurrences of five[15] verbal predicates (an average of 398 instances per predicate) selected from the OntoNotes 4.0 corpus (Weischedel et al. 2011). They followed the SemEval task 10 guidelines (Ruppenhofer et al. 2010, 2012) for the annotation and linking of null instantiations (NI), except that they distinguished between resolvable and non-resolvable NI within discourse instead of classifying them as Definite NI and Indefinite NI. They use this data to show that the performance of SRL systems, which deal with implicit arguments (or NIs), can be improved when the sparseness of the training corpus is reduced.

There also exist proposals for the automatic creation of 'artificial training data' (Silberer and Frank 2012)

[5] In AnCora corpus, the tag 'S' stands for clause.
[6] http://nlp.cs.nyu.edu/meyers/NomBank.html
[7] https://framenet.icsi.berkeley.edu/
[8] http://clic.ub.edu/corpus/ancora
[9] http://stl.recherche.univ-lille3.fr/programmesetcontrats/NOMAGE/NOMAGEenglish.html
[10] https://code.google.com/p/copenhagen-dependency-treebank/
[11] http://www.coli.uni-saarland.de/projects/semeval2010_FG/
[12] The authors also provided a version of the corpus based on PropBank/NomBank annotations.
[13] Predicates annotated: 'bid', 'sale', 'loan', 'cost', 'plan', 'investor', 'price', 'loss', 'investment' and 'fund'.
[14] Henceforth, we will refer to the Moor, Roth and Frank (2013) corpus as the MRF corpus.
[15] Predicates annotated: 'give', 'put', 'leave', 'bring' and 'pay'.

to address the sparse data problem and the scarcity of resources annotated with implicit arguments. For instance, Silberer and Frank (2012) propose a technique for the heuristic acquisition of labelled data from corpora manually annotated with coreference information and semantic roles. Basically, these authors follow an entity-based approach, in which entities are represented by their coreference chains. They artificially delete the semantic role label of the anaphoric pronouns and assign it to the closest antecedent in the coreference chain of these pronouns. Roth and Frank (2013) propose a heuristic method for acquiring a dataset of implicit arguments and their discourse antecedents, which exploits aligned predicate argument structures from pairs of comparable texts. They compare the argument structures of both predicates searching for the explicit arguments in one predicate argument structure that has been unrealized (implicit) in the other. Once the implicit arguments are identified, they link them to an antecedent taking into account the cross-document coreference chain of its explicit counterpart.

Table 1 Corpora annotated with implicit arguments

Corpus	Source	Types	Tokens	TTRatio	Predicates	Senses	AScheme	Process	Language
SemEval-10 Ruppenhofer et al. (2010, 2012)	A.C. Doyle works	769	3,073	3.99	deverb. nouns +verbs	Ambiguous Unambiguous	FrameNet	Manual	English
G&C Gerber and Chai (2010, 2012)	Penn TreeBank	10	1,247	124.7	deverb. nouns	Unambiguous	PropBank NomBank	Manual	English
MR&F Moor, Roth and Frank (2013)	OntoNotes 4.0	5	1,992	398.4	verbs	Ambiguous Unambiguous	FrameNet	Manual	English
S&F Silberer and Frank (2012)	OntoNotes 3.0[16]	258	12,770	49.49	deverb. nouns +verbs	Ambiguous Unambiguous	FrameNet PropBank VerbNet	Automatic	English
	ACE-2	757	58,204	76.88					
	MUC-6	654	20,140	30.79					
R&F Roth and Frank (2012, 2013)	IndIA[17]	450	698	1.55	deverb. nouns +verbs	Ambiguous Unambiguous	PropBank NomBank	Automatic	English
Iarg-AnCora	AnCora-Es	1,454	18,397	12.59	deverb. nouns	Ambiguous Unambiguous	PropBank NomBank VerbNet	Manual	Spanish

Table 1 summarizes the main characteristics of the above mentioned corpora: the name of the corpus and its corresponding reference (column 1); the source from which they have been created, most of them were built from existing annotated corpora (column 2); the size, specifying the types and tokens of the predicates analyzed with implicit arguments (columns 3 and 4) and their token/type ratio (column 5); the type of predicate annotated, i.e. verbs and deverbal nouns (column 6), and whether the predicates are ambiguous or unambiguous (i.e. whether they have one or more than one sense) (column 7); the annotation scheme used (column 8); the annotation process followed (column 9), and the language (column 10).

With the exception of the datasets created artificially, the above mentioned corpora share the following basic characteristics: a) they are only annotated with core arguments, so they do not deal with adjunct arguments; b) they only mark the identity relations between the referents, that is, between the antecedent and the implicit argument instance (no bridging or part-whole relations are considered); c) the instances of implicit arguments are linked to all mentions of the referents and, therefore to the coreference chain of

[16] OntoNotes 3.0 (Hovy et al. 2006), ACE-2 (Mitchell et al. 2003) and MUC-6 (Chinchor and Sundheim 2003).

[17] The IndIA (Inducing Implicit Arguments) corpus by Roth and Frank, consists of several datasets initially derived from a set of automatically extracted pairs of comparable texts from the English Gigawords Fifth Edition corpus (Parker et al. 2011), comprising pairwise documents that are predicted to be about the same events and entities. The dataset referred to the table contains 698 instances of implicit arguments and discourse antecedents that were automatically extracted from comparable texts in this initial dataset, as described in Roth and Frank (2012). The dataset includes 379 different predicates and 450 different arguments.

mentions[18]; d) they were all manually annotated; and, e) their size, especially in terms of tokens, is rather small. These characteristics are also shared by the Spanish Iarg-AnCora corpus presented in this article. But, in contrast to the English corpora, Iarg-AnCora has an extended coverage (18,397 occurrences corresponding to 1,454 different types, average of 0.64 implicit arguments per predicate) and, unlike the G&C corpus, all of the deverbal nouns are analyzed, not only a small subset of the unambiguous ones. The Spanish corpus differs from the MR&F corpus in that they annotated the implicit arguments of specific verb predicates. The SemeEval-2010 and G&C corpora are both built for specific tasks, while Iarg-AnCora and the other English corpora could be used as a reference corpus for the annotation of implicit arguments and for the linguistic analysis of this kind of phenomena. The primary goal of Iarg-AnCora is not to be a corpus for a specific semantic task, although it could obviously be used for that purpose. In fact, a subset of the Iarg-AnCora was used as a training and test corpus for creating LIARc (Peris et al. 2013) (See section 6).

We will now briefly present the systems recently developed to automatically detect and classify implicit arguments, which use as training corpora those described above. Most of them deal with English and can be split in two groups. On the one hand are those systems related to the SemEval-2010 Task 10 (Ruppenhofer et al. 2010), and concretely, those that tackled the NI resolution subtask. We include in this group the two participating systems, Semafor (Chen et al. 2010) and Venses++ (Tonelli and Delmonte 2010), and those systems that use the same data set and evaluation measures used in this subtask (Silberer and Frank 2012; Laparra and Rigau 2012; Ruppenhofer, Gorinski and Sporleder 2011; Tonelli and Delmonte 2011, and Wang et al. 2013). All these systems identify implicit arguments for different English predicates (verbs and nouns), following the typology of implicit arguments proposed in Fillmore (1986) and Fillmore and Baker (2001) and the FrameNet annotation scheme (Baker, Fillmore and Lowe 1998). These systems use different approaches to resolve the binding of NIs. Silberer and Frank (2012) approach the problem as a CR task; Tonelli and Delmonte (2011), Semafor, and Venses++ approach it as an extension of SRL systems, while Ruppenhofer, Gorinski and Spoleder (2011) and Laparra and Rigau (2012) adopt a mixed approach to carry out the task combining both strategies. These systems can use supervised ML techniques such as Silberer and Frank (2012) and Semafor, or they can be based on hand written rulesets that use different type of information (Ruppenhofer, Gorinski and Spoleder (2011); Laparra and Rigau (2012); Tonelli and Delmonte 2010 and 2011). It is also worth noting that all systems except Chen et al. (2010) (which works in parallel) deal with the problem sequentially, that is, by breaking down the task into different subtasks. Laparra and Rigau (2013) base their approach on the discourse coherence of predicates. Roth and Frank (2013) use as the core of their approach a dataset of automatically aligned predicate pairs released by Roth and Frank (2012).

On the other hand, Gerber and Chai (2010, 2012) developed a parallel supervised ML feature-based model to detect the core implicit arguments of English deverbal nominalizations, which uses G&C corpus described above. More detailed information, and some improvements can be found in Gerber (2011).

In section 6, the LIARc classifier (Peris et al. 2013), the only system dealing with the implicit arguments of deverbal nouns in Spanish, is described in more detail. This classifier uses Iarg-AnCora as a training and test corpus, and it is based on the experiments carried out by Gerber and Chai (2010, 2012).

4 Annotation of implicit arguments

In this section, we describe the annotation of the Spanish AnCora corpus with the implicit arguments of deverbal nouns. The annotation process involved two subtasks that were carried out manually. The first subtask consisted of detecting the missing core implicit arguments whose semantic interpretation depends on the linguistic context. This task also involved the assignment of the argument position -iarg0, iarg1, iarg2, etc.- and its corresponding thematic role -agent, cause, patient, among others. The second subtask consisted of linking the implicit arguments to discourse entities.

The syntactic constituents that can be annotated as antecedents of implicit arguments are: *sn* (NP), *grup.nom* (nominal group in a conjoined NP), *relatiu* (relative pronoun) and *S* (clause), that is, those

[18] Note that a coreference chain may consist of only one mention, that is, a singleton.

constituents that can be discourse entities[19]. A discourse entity can consist of only one mention -that is, a singleton- or can be a coreference chain, which consists of different types of mentions that point to the same referent (Recasens and Vila 2010). The instances of implicit arguments are linked to all mentions of the referents and, therefore, to the coreference chain of mentions. For instance, in (8) the entity *los pasajeros* is a singleton ("singleton1") because it has only one mention in this document, whereas entity 4 consists of a coreference chain of three mentions (entity4="la pasarela","la que","la pasarela"), which are underlined in the example.

(8) [Un avión de [Spanair]<entity="entity1"> <entity="entity2"> despegó ayer del [aeropuerto de Barajas]<entity="entity3"> sin esperar la retirada de [la pasarela]**<entity="entity4">** por [la que]**<entity="entity4">** acceden [los pasajeros] <entity="singleton1">. Según [Spanair]<entity="entity1">, ni [el avión]<entity="entity2"> ni [la pasarela]**<entity="entity4">** sufrieron daños.

'[A [Spanair]<entity="entity1"> plane]<entity="entity2"> took off from [Barajas airport]<entity="entity3"> yesterday without waiting for the withdrawal of [the boarding bridge]**<entity="entity4">** through [which]**<entity="entity4">** [the passengers]<entity="singleton1"> board. According to [Spanair]<entity="entity1">, neither [the airplane]<entity="entity2"> nor [the boarding bridge]**<entity="entity4">**] suffered damage'.

It is worth noting that we only marked the identity relations between the antecedent and the implicit argument instances, therefore, bridging or part-whole relations were not considered. In fact, these relations were also omitted in the annotation of AnCora corpus with coreference relations.

We use the verbal and nominal lexicons -AnCora-Verb (Aparicio, Taulé and Martí 2008) and AnCora-Nom (Peris and Taulé 2011)- as lexical resources to obtain the information about the possible implicit arguments for each predicate. The arguments to be localized in the local discursive context, and to be annotated, are those specified in the nominal or verbal lexical entries and not explicitly realized in the NP.

We use both lexicons because the verbal one is larger than the nominal one. Moreover, the verbal lexicon was manually created, whereas the nominal lexicon was automatically obtained from the annotation of the explicit arguments of nominalizations in the AnCora corpus. Therefore, only explicit arguments are represented in the AnCora-Nom lexicon. This is why we also need to consult the verbal lexicon to obtain the information missing in AnCora-Nom.

In order to ensure the quality and the consistency of the annotated data, an inter-annotator agreement test was conducted on a subsample of 200 deverbal noun tokens (out of the 18,397 tokens finally annotated) and 500 implicit arguments were revised.

In the following subsections, we introduce the annotation scheme used for the annotation of implicit arguments (subsection 4.1), then we describe the annotation process (subsection 4.2), the linguistic resources (subsection 4.3) and the annotation tool (subsection 4.4) and, finally, we provide details about the inter-annotator agreement test (4.5).

4.1 Annotation scheme

The annotation scheme used for tagging the implicit arguments is the same as the one followed to annotate the explicit arguments of deverbal nouns (Peris and Taulé 2011) and the argument structure of verbs in AnCora (Taulé, Martí and Recasens 2008), which was in turn based on PropBank/NomBank for argument annotation and VerbNet (Kipper 2006) for the annotation of thematic roles. In this way, we ensure the consistency of the annotation of arguments of different predicates -nouns and verbs-, as well as the compatibility of Spanish and English resources.

For the sake of simplicity, we use the symbol [Ø] to indicate the presence of an implicit argument. We use the *iargn="r:entity$_x$"* tag abbreviation to identify implicit arguments and to differentiate them from

[19] Possessive pronouns and determiners can also be discourse entities, but they do not tend to be implicit arguments of deverbal nouns since they usually appear explicitly inside of the NP headed by the nominalization. For instance, *Esto permitirá al banco sanear sus cuentas, que es condición básica para continuar con su privatización*, 'This will enable the bank to consolidate its accounts, which is a basic condition for its privatization'. In this example, the possessive determiner *su* ('its') is the explicit argument, with the thematic role theme, of the deverbal noun *privatización* ('privatization').

explicit arguments (*argn="r"* tag) (Gerber and Chai 2010, 2012). In this tag, the letter *i* identifies implicit arguments and *n* indicates the argument position (from 0 to 4). The *r* attribute tag is used to indicate the thematic role and the $entity_x$ attribute tag indicates the discourse entity to which it is linked (*x* indicates the entity number). The list of thematic roles includes 20 different labels based on VerbNet proposal. The combination of the five argument position labels (iarg0, iarg1, iarg2, iarg3, iarg4) with the different thematic roles results in a total of 23 possible semantic tags[20] (Table 2).

In order to link an implicit argument to its corresponding discourse entity -a singleton or a coreference chain-, we take into account the coreference information tagged in the AnCora corpus. Therefore, we follow the same annotation scheme used in the coreference annotation (Recasens and Marti 2010), which was in turn based on the general criteria of the MATE scheme (Poesio 2004, Poesio and Artstein 2005).

The link is established by anchoring the implicit argument (*iargn*) to the corresponding discourse entity, concretely by the attribute *entity*. The possible values of *entity* can be a 'singleton' (identifying discourse entities with only one mention) (11), an 'entity' (identifying the mentions of a coreference chain) (9) or the combination of two discourse entities (either singletons or coreference chains, for instance '$entity_n$+$entity_n$' or '$singleton_n$+$entity_n$') (10). Each mention has an entity number ('$entity_n$' or '$singleton_n$') assigned to it, and all the mentions of a coreference chain share the same entity number. Each mention also has its associated *entityref,* an attribute for indicating whether the mention is referential or not. This attribute has five possible values: 'ne' refers to a named entity mention; 'nne' stands for a non-named entity mention; 'spec' basically refers to anaphoric pronouns; 'lex' indicates non-referential mentions that are part of an idiom; and, finally, no *entityref* stands for the mentions which are not referential. In the case of coreference chains, the attribute *coreftype*="ident" indicates an identity relation between the antecedent and the implicit argument, the only coreferential relation annotated in AnCora.

Table 2 Values of the attributes *entity*, *iarg* and *r*

Attribute <entity> value	**Attribute <iarg> value**	**Attribute <r> value**
$entity_n$ $singleton_n$ $entity_n$ + $entity_n$ $singleton_n$ + $singleton_n$ $entity_n$ + $singleton_n$ Syntactic tags that can be antecedents of an implicit argument: sn, S, grup.nom, relatiu[21]	iarg0	agt (agent) cau (cause) exp (experiencer) src (source)
	iarg1	loc (locative) pat (patient) tem (theme)
	iarg2	atr (attribute) ben (beneficiary) cot (co-theme) efi (final state) ein (initial state) exp (experiencer) ext (extension) loc (locative) tem (theme)
	iarg3	ben (beneficiary) ori (origin) ein (initial state) ins (instrument) loc (locative)
	iarg4	des (goal) efi (final state)

[20] Not all the combinations of argument position and thematic roles are valid semantic tags.
[21] See the introduction to section 4 for a detailed explanation of these tags.

(9) [[La alcadesa]<entity="entity24" coreftype="ident" entityref="ne" ne="person"> lanzó [duras **críticas [Ø]**<iarg0="agt:entity24">] [contra los dirigentes deportivos que no defendieron su triunfo]<arg1="pat">]sn]sentence.
'[[The Mayor]<entity="entity24" coreftype="ident" entityref="ne" ne="person"> launched [harsh criticisms **[Ø]**<iarg0="agt:entity24">] [against the sport leaders, who did not defend her victory] <arg1="pat">]sn]sentence'.

In sentence (9), the agent implicit argument of the deverbal noun *críticas* ('criticisms') (iarg0="agt:entity$_{24}$") is linked to the discourse entity *la alcaldesa* ('the Mayor') by the attribute *entity* (entity="entity$_{24}$"), which is a named entity referential mention (entityref="ne"). Since the entity is a mention from a coreference chain, the *coreftype* attribute indicates that an identity relation (coreftype="ident") is established with the other mentions in the chain.

(10) **[**Según Spanair, [ni [el avión]<entity="entity2" coreftype="ident"> ni [la pasarela]<entity="entity4" coreftype="ident">]sn_coord sufrieron [**daños [Ø]**<iarg1="tem:entity2+entity4">]sn]sentence.
'[According to Spanair, [neither [the airplane]<entity="entity2" coreftype="ident"> nor [the boarding bridge]<entity="entity4" coreftype="ident">]sn_coord suffered [**damage [Ø]**<iarg1="tem:entity2+entity4">]sn]sentence'.

(11) La construcción del Fòrum exigirá [[la demolición de los 80 pisos]<entity="entity29" coreftype="ident"> y [la colocación de los vecinos en nuevas viviendas]<entity="singleton4">]sn_coord. [La **operación [Ø]**<iarg1="pat:singleton4+entity29">]sn no se acabará antes de dos años.
'The construction of the Forum will require [[the demolition of 80 flats] <entity="entity29" coreftype="ident"> and [the rehousing of the neighbours in new houses]<entity="singleton4">]sn_coord. [The operation [Ø]<iarg1="pat:singleton4+entity29">]sn will take at least two years.'

In sentence (10), the theme implicit argument of the deverbal noun *daños* ('damages') is linked to two different discourse entities, *el avión* ('the airplane') and *la pasarela* ('boarding bridge') (iarg1="tem:entity2+entity4"), which are part of a coordinated NP (sn_coord). In sentence (11), the patient implicit argument of the deverbal noun *operación* ('operation') is linked to a singleton entity and to an entity which is part of a coreference chain, (iarg1="pat:singleton4+entity29"). The combination of different discourse entities is often due to the presence of coordinated NPs, as is shown in the above examples.

4.2 Annotation process

The steps we followed in the annotation process were: a) first, to identify the missing core arguments (*iargn=r:entity$_x$*), taking into account the information contained in the AnCora lexicons, and to assign their argument position and the corresponding thematic role; b) second, to find the discourse entity (*entity* or *singleton*) in the discursive context, that is, the antecedent to which to link the implicit argument. Singletons were not annotated in AnCora-Es, but in the Iarg-AnCora corpus they were tagged when they were the antecedents of the implicit argument of a deverbal noun. If it was not possible to find an antecedent, the implicit argument remained unresolved and no specific tag was associated. It is worth noting that, in contrast to Gerber and Chai (2010, 2012), we can link an implicit argument to mentions appearing not only within the sentence containing the deverbal noun and within preceding sentences, but also in subsequent sentences. Unlike Ruppenhoffer et al. (2010, 2012), we can also link the arguments to singletons and not only to mentions in coreference chains. Singletons are less likely to be antecedents of implicit arguments (23% in the corpus) than entities in coreference chains (76.69%), but they cannot be ignored.

4.3 Linguistic resources

The main linguistic resource used for building Iarg-AnCora is the AnCora-Es corpus, a Spanish multi-layered annotated corpus, which consists of 400,000 words derived from newspaper and newswire articles.[22] This corpus was morphologically tagged (with PoS and lemma information), syntactically parsed (with constituents and functions), semantically annotated (with the argument structure of verbs and deverbal nominalizations, WordNet[23] nominal senses and named entities) and, finally, annotated at the

[22] 200,000 words were extracted from the Spanish *El Periódico* newspaper (http://www.elperiodico.com/es/) and the other 200,000 words from the EFE newswire agency (http://www.efe.es), spanning from January to December 2000.
[23] We used Spanish WordNet in the Multilingual Central Repository (MCR), which is linked to Princeton WordNet (Gonzalez-Agirre, Laparra and Rigau 2012), http://adimen.si.ehu.es/web/MCR.

discourse level (with coreference information). All of these annotated layers were manually validated in order to ensure the quality of the final resource. The annotation of the verbal and nominal argument structures only dealt with the arguments explicitly realized, which is why we have enriched the corpus with the implicit arguments of deverbal nouns. As for the coreference information, it includes the coreference links between pronouns (including elliptical subjects and clitics),[24] full NPs (including proper nouns) and discourse segments (one or more contiguous sentences), as well as the type of coreference relation established -identity, discourse deixis and predicative relations- (Recasens and Martí 2010).

AnCora-Verb[25] is a Spanish lexicon consisting of 2,830 verbal entries, which correspond to the verbs appearing in the corpus. A verb can have different senses and each sense can have different syntactic-semantic frames depending on the diathesis alternations in which it can participate.[26] Each frame provides the mapping between a syntactic function and its constituent, argument position and thematic role, as well as the semantic class to which it belongs. Relevant examples of uses for each frame extracted from the corpus are also provided. Currently, there are 24 different semantic classes (Taulé, Martí and Borrega 2011)[27], which are based on the proposal of Levin (1993). Figure 1 shows the information associated with the entry of *criticar* ('to criticize') in AnCora-Verb. The first sense of *criticar* (<sense id="1">) has two frames with their corresponding semantic classes associated with them: the first belongs to the transitive-agentive-patient semantic class (lss="A21"), and the second to the unaccusative-passive-transitive semantic class (lss="B22"), which corresponds to the passive alternation. In the transitive frame, the subject (suj) maps to the first argument (arg0) with the thematic role of agent (agt), whereas the object (cd) corresponds to the second argument (arg1) with the thematic role of patient (pat). In the passive frame, there is an argument crossing: the affected object appears as subject (suj) and maps to the second argument (arg1) with the thematic role of patient (pat); and the agent maps the first argument (arg0) with the agent complement (cag), which is syntactically realized by a prepositional phrase (sp).

lemma="**criticar**"
type="**verb**"
sense id="**1**"
frame_type="**transitive-agentive-patient**"
lss="**A21**"
argument="**arg0**" function="**suj**" thematicrole="**agt**"
argument="**arg1**" function="**cd**" thematicrole="**pat**"
example= El secretario **criticó** que la temporalidad de los contratos impide la caída del paro.
('The secretary criticized the fact that temporary employment contracts prevent unemployment from falling')
frame_type="**unaccusative-passive-transitive**"
lss="**B22**"
argument="**arg1**" function="**suj**" thematicrole="**pat**"
argument="**arg0**" fuction="**cag**" thematicrole="**agt**" constituent type="sp" "preposition="por"
...

Fig. 1 Lexical entry of *criticar* ('criticize')

AnCora-Nom[28] is a lexicon of Spanish deverbal nominalizations consisting of 1,658 entries, which corresponds to the deverbal nouns appearing in the corpus. Each sense of a deverbal noun has an associated denotation type (i.e., event, result, and underspecified), an assigned WordNet synset. The mapping of nominal complements with arguments and the corresponding thematic roles is also annotated. This mapping is established taking into account the syntactic and semantic information of the verb base from which the nominalization is derived and is represented in AnCora-Verb. The AnCora-Nom lexical entries are linked to their corresponding verbal lexical entries in AnCora-Verb.

Figure 2 shows that *crítica* ('criticism') is a deverbal noun (origin="deverbal" type="noun") linked to the first sense of the *criticar* verbal entry (originlink="verb.criticar.1"), with which it shares the same argument structure -that is, the first argument (arg0) with the thematic role of agent (agt) and the second (arg1) with the thematic role of patient (pat), which is syntactically realized with a prepositional phrase constituent (sp).

[24] Spanish is a pro-drop language, therefore, pronominal subjects can be omitted. The object personal pronouns often appear as clitic forms and can be adjoined to the verb.
[25] AnCora-Verb-Es lexicon is available at: http://clic.ub.edu/corpus/ancoraverb_es
[26] AnCora-Verb contains 3,934 different senses and 5,117 syntactic-semantic frames in total.
[27] http://clic.ub.edu/corpus/webfm_send/50
[28] http://clic.ub.edu/corpus/ancoranom_es

lemma="**crítica**"
origin="**deverbal**"
type="**noun**"
sense id="**1**" denotation="**result**" originlemma="**criticar**"
originlink="**verb.criticar.1**" synset="**16:05032854**"
argument="**arg0**" thematicrole="**agt**" constituent type="**sp**" preposition="de"
argument="**arg1**" thematicrole="**pat"** constituent type="**sp**" preposition="a"|"contra"|"sobre"
constituent postype="article" type="determiner"
example= La alcadesa lanzó duras críticas contra los dirigentes deportivos que no defendieron su triunfo.
('The Mayor launched harsh criticisms against the sport leaders, who did not defend her victory.')

Fig. 2 Lexical entry of *crítica* ('criticism')

Since the lexicons are generated from the annotated corpus, we took into account the argument structure information declared in both lexical resources to find the possible implicit arguments of each nominal predicate, i.e. those specified in the nominal or verbal lexicons but explicitly unrealized in the local context of the deverbal noun. For instance, in sentence (12) the patient argument (arg1="pat") is the only argument explicitly realized in the NP. We use the nominal and verbal lexicons to infer that there is an agent argument (arg0="agt") to locate, which, in fact, is implicitly understood (*la alcaldesa*, 'the Mayor') recovered from the same sentence but outside of the NP headed by the nominalization.

(12) La alcadesa_{**<entity="entity1">**} lanzó [duras **críticas** [Ø]_{**<iarg0="agt:entity1">**} [contra los dirigentes deportivos que no defendieron su triunfo]_{<arg1="pat">}]_{NP}.
'The Mayor_{**<entity="entity1">**} launched [harsh **criticisms** [Ø]_{**<iarg0="agt:entity1">**}] [against the sport leaders, who did not defend her victory]'.

4.4 AnCoraPipe annotation tool

In order to minimize errors in the annotation process and make the annotator's work easier, we used AnCoraPipe[29] (Bertran et al. 2011) to annotate the implicit arguments. This is an environment that enables the creation, editing and analysis of corpora and lexicons. Concretely, the edition process allows for the annotation of corpora using different linguistic interfaces, which are specific for each layer of linguistic analysis. For instance, it was also used for the annotation of the argument structure, named entities and coreference relations in the AnCora corpus. The interfaces integrated in AnCoraPipe were developed with the participation of linguists with the aim of being user-friendly and user-oriented. This resulted in a tool designed for operational simplicity through the minimization of the mouse clicks required to perform operations, the highlighting of the relevant nodes to be annotated and access to specific windows (panels) that allow us to consult, for instance, the AnCora lexicons, but also external lexical resources, such as the Multilingual Central Repository, which can be useful for the semantic annotation of corpora. In addition, it allows the different annotators to work simultaneously with the same version of the corpus. AnCoraPipe is implemented as a plug-in in the Eclipse[30] development platform. Eclipse facilitates the integrated management and collaborative building of linguistic resources using the Subversion (SVN) version control system to update the remote copies. In AnCoraPipe, the corpora texts and the lexical entries are XML documents with UTF-8 encoding.

Although AnCoraPipe was built for supporting the building and maintenance of AnCora resources (lexicons and corpora for Spanish and Catalan languages), the tool can also be configured for working with other languages[31] and purposes.

[29] AnCoraPipe is freely available, to access contact amarti@ub.edu.
[30] http://www.eclipse.org/
[31] AnCoraPipe has been used for the treatment of corpora in the Amazighe, Latin and Cyrillic alphabets.

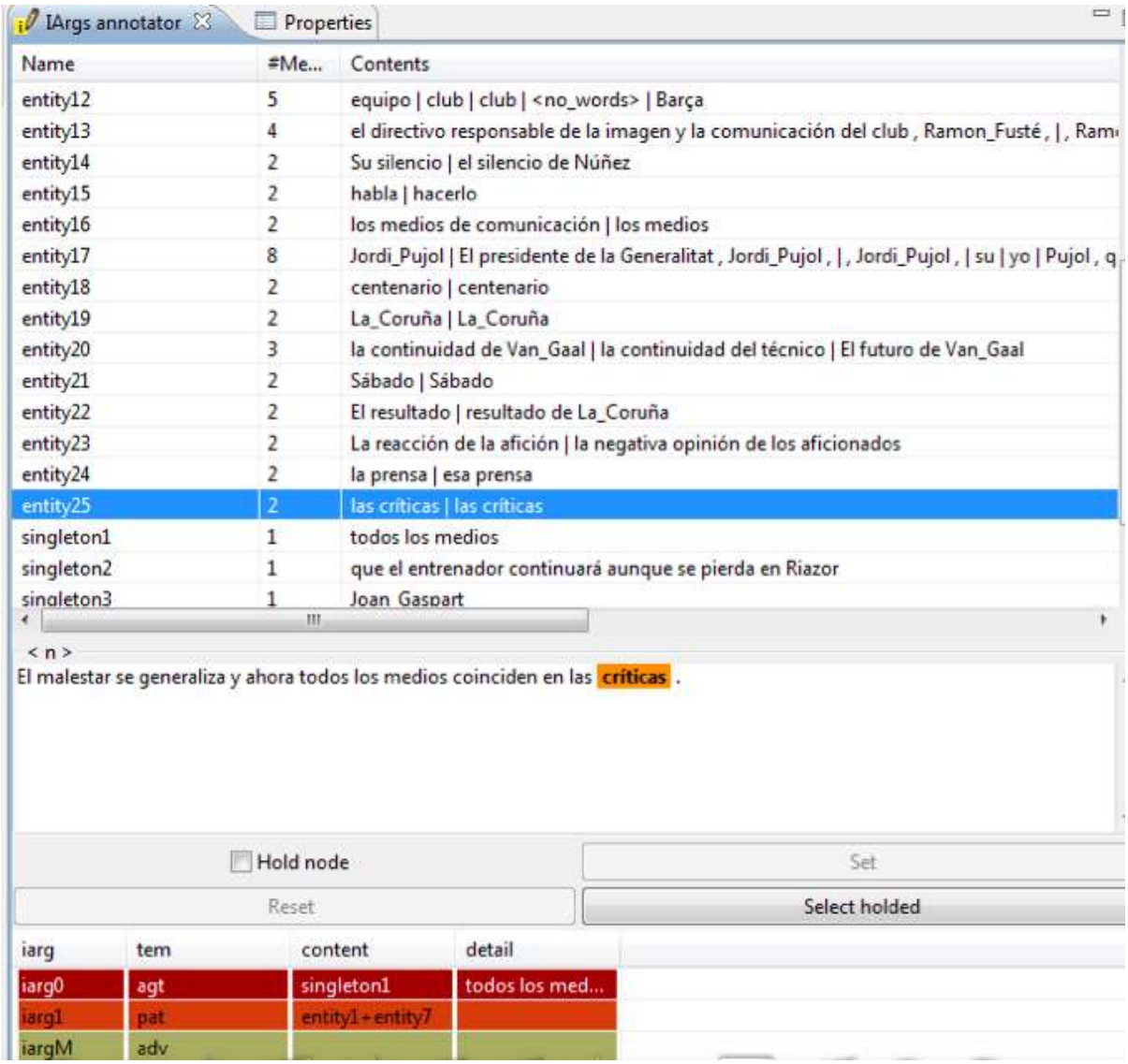

Fig. 3 A screenshot of the implicit argument annotation tool

Iarg-Annotator, a specialized user-oriented interface for the annotation of implicit arguments, specially designed to carry out this task (Figure 3), is integrated in AnCoraPipe. All of the entities that appear in the document are listed in the Iarg-Annotator panel, and the candidates for implicit arguments to be identified appear at the bottom of the panel (iarg0="agt" and iarg1="pat", in this example). The annotator has to select the correct discourse entities manually from the list of entities available in the document: in the example, the singleton$_1$ (*todos los medios*, 'all media') for iarg0="agt" and the combination of two discourse entities, entity$_1$+entity$_7$, for iarg1="pat" (entity$_1$ corresponds to *el entrenador|Van_ Gaal*, 'the coach|Van_Gaal' and entity$_7$ corresponds to *el president del club|Núñez|...*, 'the president of the club|Núñez|…')[32]. The panel also displays the sentence with the deverbal noun highlighted (*críticas*, 'criticisms').

[32] For reasons of space, Figure 3 only shows the discourse entities starting from entity$_{12}$.

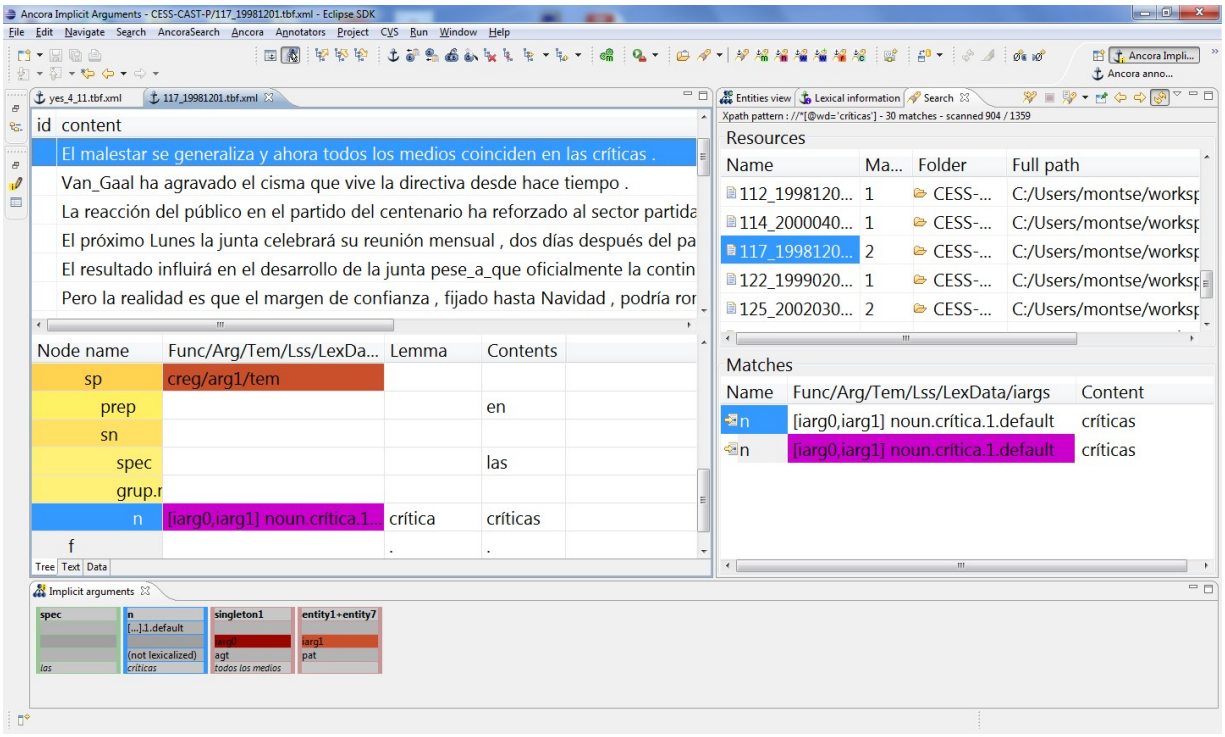

Fig. 4 A screenshot of the implicit argument annotation tool

The panel on the left in Figure 4 contains two windows: the upper one displays the text separated into paragraphs with the sentence containing the deverbal noun highlighted; the window below shows the tree structure (first column) with associated syntactic and semantic information, i.e. the syntactic function and the argument position and thematic role of the explicit arguments, as well as the possible implicit arguments to be identified (second column), corresponding to the deverbal noun (iarg0="agt" and iarg1="pat", in this example); the third and fourth columns show the lemma and the words contained in each syntactic node respectively. The panel on the right displays all the Iarg-AnCora files containing occurrences of the deverbal noun to be analyzed along with the file under revision and its specific occurrence highlighted (the noun *críticas* appears twice in the text, but in this example we are annotating the second occurrence, which is highlighted). Finally, a summary of the implicit arguments, which are identified and associated to their corresponding discourse entities, are shown at the bottom of the screen: the implicit arguments of the noun *críticas* (second column) are iarg0="agt:singleton1" (third column) and iarg1="pat:entity$_1$+entity$_7$" (fourth column).

Therefore, the annotator has to select the correct discourse entity for each implicit argument, taking into account the discourse context and the information specified in the lexicons, which can also be consulted from AnCoraPipe.

One of the benefits of using this annotation tool is that the annotators have all the necessary information for the annotation available on the same screen[33]. The panels can be laid out in the graphic space and opened and closed independently, depending on the preferences of the annotator, making the annotation process easier.

4.5 Interannotator Agreement Test

The manual annotation of Iarg-AnCora was carried out by three trained graduate students in linguistics, who also participated in the annotation of the explicit arguments of verbs and deverbal nouns in AnCora. Therefore, they were already familiar with the annotation scheme and tool to be used. In fact, they needed only minimal instructions to use the new panel specially created for this task. Once they were familiar with the annotation guidelines of implicit arguments (Peris and Taulé 2013), an Inter-Annotator

[33] We have split the panels in two figures in order to better visualize their content.

Agreement test was conducted to ensure the consistency and quality of the Iarg-AnCora annotation. The analysis of disagreements enabled us to check whether the annotators had clearly understood the task, to detect the most problematic aspects in the annotation, to resolve them and to improve the annotation guidelines in order to obtain better results.

The data consisted of a subsample of 200 deverbal noun tokens corresponding to 8 unambiguous lemmas -*actuación* ('actuation'), *comunicado* ('communication'), *daño* ('harm'), *empate* ('draw'), *negocio* ('business'), *opción* ('option'), *propuesta* ('proposal') and *viaje* ('travel')-, which appear in 88 different documents (files). We selected unambiguous lemmas because the most difficult task for the annotators was to identify the correct discourse entity to which the implicit argument had to be linked. Unambiguous lemmas are monosemous nouns, which do not present problems of ambiguity in the selection of argument position and thematic role. Since each deverbal noun can have more than one implicit argument to be identified, a total of 500 possible implicit arguments were reviewed, some of them were assigned to a discourse entity and other candidates were unresolved. Each annotator tagged the documents separately.

We did not compute the agreement in the assignment of thematic roles because, on the one hand, the annotators had prior annotation experience and they had participated in the annotation of the explicit arguments of verbs and nominalizations[34] and, on the other hand, because the argument positions and thematic roles are specified in the lexicons and there was little room for error.

Table 3 Inter-Annotator Agreement: pairwise and total agreement

Pairwise agreement					**Total agreement**
Annotator pairs	**A-B**	**A-C**	**B-C**	**Average**	**A-B-C**
Observed agreement	79%	80%	84%	81%	71%
Cohen's Kappa	0.54	0.56	0.65	0.58	0.39

In Table 3, we present the pairwise and total agreement percentages obtained. Columns show the result for each pair of annotators (pairwise agreement) and between all the annotators (total agreement). The rows show the observed agreement and Cohen's kappa coefficient (as in Gerber and Chai (2010) in order to compare the results obtained). There is agreement when all the annotators link the implicit argument to the same discourse entity, or when they link it to at least one of the discourse entities if there is more than one. The average pairwise result obtained among the three pairs of annotators was 81% of observed agreement (0.58 kappa). The total agreement obtained was 71% (0.39 kappa). The results obtained show a moderate agreement that confirms the complexity of the task.
A direct comparison with agreements obtained by other authors is difficult because of the differences in language and predicates. The most fair comparison could be made between our pairwise average kappa (0.58) and the one reported by G&C (0.67).

The main source of disagreement (85% of cases) is due to missed links, that is, one of the annotators links an implicit argument to a discourse entity and the others do not, or one annotator does not recognize the link that the other two have identified. Moreover, in 75% of cases, the source of this type of disagreement is the least experienced annotator. The remaining 15% corresponds mainly to disagreements due to the three following reasons:

a) Different interpretations of the antecedent, especially when a singleton is involved:

(13) [Pasqual_Maragall]$_{\text{<entity="entity2" coreftype="ident">}}$ acudirá a la reunión con un paquete de [*propuestas* [de [diálogo]$_{\text{<entity="singleton1">}}$]$_{\text{<arg1="pat">}}$]$_{\text{sn}}$ […]. [Pasqual_Maragall]$_{\text{<entity="entity2" coreftype="ident">}}$ se centrará en exponer [sus$_{\text{arg0="agt"}}$ **propuestas**$_{\text{iarg1="pat:singleton1">}}$]$_{\text{sn}}$ […].

'[Pasqual_Maragall]$_{\text{<entity="entity2" coreftype="ident">}}$ will attend the meeting with a package of [*proposals* [for [dialogue]$_{\text{<entity="singleton1">}}$]$_{\text{<arg1="pat">}}$]$_{\text{sn}}$ […]. [Pasqual_Maragall]$_{\text{<entity="entity2" coreftype="ident">}}$ will focus on presenting [their **proposals**$_{\text{<iarg0="agt:entity2" iarg1="pat:singleton1">}}$]$_{\text{sn}}$ […]. '

[34] The mean of inter-annotator agreement for the annotation of explicit arguments reached 0.75 kappa, which translated to 79.2% observed agreement.

In example (13), for instance, one annotator links the iarg1="pat" of *propuestas* (‘proposals’) to the singleton1 *diálogo* (‘dialogue’) in a preceding sentence, whereas another annotator links the argument to the PP *(de diálogo*) which includes the noun *diálogo*. Probably the fact that the PP *‘de diálogo’* is already tagged as an explicit patient argument (<arg1="pat">) of the previous noun *propuestas* influenced her decision.

b) Different interpretations when the deverbal noun is part of a multiword expression:

(14) Portavoces de [Chupa-Chups]<entity="entity2" coreftype="ident"> aseguraron que adquirir [el parque de atracciones]<entity="entity3" coreftype="ident"> a tan buen precio era una oportunidad para apostar por esta línea de [**negocios**<iarg0="agt:entity2" iarg1="pat:entity3">]sn.

'[Chupa-Chups]<entity="entity2" coreftype="ident"> spokespeople stated that the acquisition of [the amusement park]<entity="entity3" coreftype="ident"> at such a good price was an opportunity to pursue this line of [**business**<iarg0="agt:entity2" iarg1="pat:entity3">]sn'.

In example (14), one annotator links the iarg0="agt" of *negocios* (‘businesses’) to the entity2 *Chupa-Chups* (iarg0="agt:entity2") and the iarg1="pat" to the entity3 *el parque de atracciones* (‘amusement park’) (iarg1="pat:entity3"), whereas the other two annotators do not. In this example, the fact that *negocios* appears in a multiword expression may have influenced their choice.

c) Metaphorical vs. literal interpretation of the deverbal noun:

(15) Al [lobo de [la derecha]<entity="entity3" coreftype="ident">]<entity="entity2" coreftype="ident"> se le ven las orejas y las intenciones. Hay quien confunde [el viaje<iarg0="agt:entity2" iarg3="ori:entity3"> al centro<arg4="dest">]sn con una excursión dominguera.

'One can start to see [the wolf’s ears of [the right]<entity="entity3" coreftype="ident">]<entity="entity2" coreftype="ident"> and their intentions. There are people that confuse [a journey<iarg0="agt:entity2" iarg3="ori:entity3"> to the centre<arg4="dest">]sn with a Sunday outing'.

In example (15), only one annotator links the iarg0="agt" of *viaje* (‘journey’) to the entity2 *el lobo de la derecha* (‘the wolf of the right’) and the iarg3="ori" to the entity3 *la derecha* (‘the right’). A possible explanation for this disagreement is the general metaphorical sense of the whole text, where ‘the wolf of the right’ is interpreted as a right wing political party and the journey from the right to the centre as the change of ideology that this party is experiencing.

Due to the complexity of the task and the problems detected in the inter-annotator agreement test, the guidelines were revised; annotator A, who presented less agreement, was excluded from the task, and the other two annotators received more training before proceeding with the annotation of the whole corpus. The remaining annotation was not conducted in parallel.

In order to ensure the consistency in the annotation, we tagged the corpus in two stages: First, the unambiguous deverbal nouns were annotated, i.e. the monosemous nouns, which do not present problems of ambiguity in the selection of argument position and thematic role. In this stage, an expert annotator validated the annotation focusing on the linking of arguments to the discourse entities. In a second stage, the ambiguous deverbal nouns were annotated.

The annotation was carried out by lemma, that is, the same annotator tagged all the occurrences of the same (unambiguous or ambiguous) lemma, instead of annotating all the occurrences of different lemmas in the same document. We also had weekly meetings until the end of the annotation process in which difficult and doubtful cases were discussed and documented.

5 Statistics on the content of Iarg-AnCora

In this section, we present distributional statistics for the implicit arguments in Iarg-AnCora, which can be useful for linguistic analysis. From the total of 18,397 deverbal noun instances (tokens) annotated corresponding to 1,454 different lemmas (types), we highlight the following observations for Spanish.

1. Implicit arguments are more frequent than explicit arguments in nominal predicates. 83.8% of the 1,454 deverbal nouns analyzed have at least one implicit argument. The average number of implicit arguments realized among the predicates analyzed is 0.91 implicit arguments per nominal instance, while the average number of explicit arguments was 0.58. Consequently, the overall number of arguments is 1.50 (average) per nominal instance. Therefore, the annotation of implicit arguments is crucial for the semantic treatment of deverbal nominalizations and provides a gain in role coverage of 128%. These figures are much higher than those reported by Gerber and Chai (2012) for English, where a 71% relative gain in role coverage across the 1,247 annotated instances is obtained. Although these figures are not directly comparable, the difference is due to the lower number of explicit arguments for Spanish deverbal nouns (0.5 on average) compared to English (1.1 on average).

   Table 4 Average size (in tokens) of NPs and VPs in English and Spanish. 'All' refers to all the phrases, 'Top' to the highest scope phrases, and 'Base' to the basic ones

	English		**Spanish**	
	NP	**VP**	**NP**	**VP**
All	3.75	11.30	10.18	29.42
Top	4.24	13.62	12.46	55.03
Base	2.14	7.49	2.84	10.33

   In order to perform meaningful comparisons between the distributions of explicit and implicit arguments in Spanish and English, in Table 4 we present information on the average length (in tokens) of nominal (NP) and verbal (VP) phrases for both languages. We compute the figures for three cases: i) all the phrase, i.e. the number of tokens corresponding to leaves in the trees rooted by NP (respectively for VP), in the row labeled 'all', ii) only the phrases not included in other phrases of the same type, in the row labeled 'top', and iii) only the phrases not including phrases of the same type, in the row labeled 'base'. The statistics have been obtained from subsets of the Penn TreeBank and AnCora-Es corresponding to the dataset used by G&C for English and the similar dataset used for Spanish in the LIARc experiment. From the table, we can confirm our intuition that Spanish is more verbose than English: on average, Spanish base VPs are 1.3 times longer than English ones, and Spanish base NPs are 1.2 times longer than English ones. VPs are longer than NPs for both English (3.4 times longer) and Spanish (3.6 times longer).

   Table 5 contains the percentages of explicit (e-args), implicit (i-args) and non-resolvable (nr-args) noun arguments (columns) per argument position (rows). In each cell of the table, as well as the absolute count of the instances[35] found, we include the relative contributions in percentage for the argument type (in rows, tagged with ➔) and for the argument position (in columns, tagged with ➘). The rows show the number and percentage of explicit, implicit and non-resolvable cases for each argument position (arg0-arg4). For instance, arg0 is realized in 19.28% of cases as an explicit argument, 53.29% as an implicit argument and 27.43% as a non-resolvable argument. The columns show the distribution for argument position for each type of argument (explicit, implicit and non-resolvable arguments), for instance, 60.54% of explicit arguments are arg1 and 28.18% are arg0; and 49.49% of implicit arguments are arg0 and 37.25% are arg1. These figures are in bold in the table for the sake of clarity.

[35] Instances stand for the number of occurrences of argument types found in the corpus.

Table 5 Percentages of explicit, implicit and non-resolvable arguments per argument position

e-args		i-args		nr-args		total
arg0	2,995➔**19.28**% 🡻 **28.18**%	**iarg0**	8,277➔**53.29**% 🡻 **49.49**%	**nrarg0**	4,260➔**27.43**% 🡻 26.3%	15,532➔100% 🡻 35.66%
arg1	6,434➔37.23% 🡻 **60.54**%	**iarg1**	6,230➔36.05% 🡻 **37.25**%	**nrarg1**	4,617➔26.72% 🡻 28.51%	17,281➔100% 🡻 39.68%
arg2	1,044➔11.94% 🡻 9.82%	**iarg2**	1,924➔22.01% 🡻 11.5%	**nrarg2**	5,775➔66.05% 🡻 35.65%	8,743➔100% 🡻 20.08%
arg3	41➔4.06% 🡻 0.39%	**iarg3**	128➔12.67% 🡻 0.77%	**nrarg3**	841➔83.27% 🡻 5.19%	1,010➔100% 🡻 2.32%
arg4	114➔11.59% 🡻 1.07%	**iarg4**	166➔16.87% 🡻 0.99%	**nrarg4**	704➔71.54% 🡻 4.35%	984➔100% 🡻 2.26%
total	10,628➔24.4% 🡻 100%		16,725➔38.4% 🡻 100%		16,197➔37.19% 🡻 100%	43,550➔100% 🡻 100%

From the total number of possible candidate arguments for annotation (43,550), 24.4% are realized explicitly, 38.4% are implicit arguments and 37.1% are non-resolvable arguments (last row of Table 4). Therefore, these results show that the arguments in deverbal nominalizations are realized more implicitly than explicitly. The non-resolvable arguments (37%) are those that cannot be recovered from the linguistic context, but could probably be recovered from the extralinguistic context. It is worth noting that most non-resolvable arguments are arg2 (66.05%), arg3 (83.27%) and arg4 (71.54%), which correspond to the semantic roles of origin, goal, locative, instrument, initial and final state, that is, more optional semantic roles. Among the core arguments, it is worth noting that those arguments closest to the predicate (arg0, arg1) are more frequently realized (explicitly and implicitly) –72.57% and 73.28% respectively (adding columns 2 and 4)- than the remaining arguments: arg2 (33.95%), arg3 (16.73%) and arg4 (28.46%).

It is also interesting to highlight that arg0 is more frequently realized as an implicit argument (49.49%) than as an explicit argument (28.18%), whereas arg1 is more frequently realized as an explicit argument (60.54%) than as an implicit argument (37.25%). This is probably due to the pro-drop nature of Spanish. On the other hand, arg1 is the argument position that is closest to the predicate and it is necessary to complete the meaning of the deverbal noun.

In Table 6 we present the distribution of implicit arguments per argument position for English (from G&C dataset) and Spanish (from AnCora). In Table 6 we can see that iarg0 and iarg1 are the implicit arguments that appear most frequently in both languages. However, iarg1 is more implicitly realized in Spanish than in English, while the opposite holds true for iarg0 in English.

Table 6 Percentages of implicit arguments per argument position for English (G&C) and Spanish (AnCora)

Argument	English (G&C)	Spanish (AnCora)
iarg0	35.23%	28.18%
iarg1	33.07%	60.54%
iarg2	23.01%	9.82%
iarg3	8.14%	0.39%
iarg4	0.55%	1.07%

Table 7 shows the number of deverbal noun instances distributed according to their number (0-5) of explicit and implicit arguments. For instance, there are 2,296 instances without any explicit or implicit argument realized, 3,127 instances with only one explicit argument, 3,251 instances with only one implicit argument, and 4,009 instances with one explicit argument and one implicit

argument. The last row contains the total percentages of implicit arguments and the last column shows the total percentages of explicit arguments. It is worth noting that in 35.03% of cases all the arguments are explicit and in 49.57% of cases all the arguments are implicit, a difference of 15 points. In contrast, the percentage of instances with more than two explicit arguments realized is smaller than 7%, and the percentage with more than two implicit arguments is higher than 23%.
This can be explained by the proper function of nominalizations which focus more on the event expressed than on the participants in the event, which have usually been presented previously in the discourse. This can also be explained by the fact that NPs are constituents that do not tend to be as long as VPs, therefore, the number of explicit arguments in NPs is generally lower than in VPs, as has been mentioned before regarding Table 4.

Table 7 Number of deverbal noun instances with and without explicit and implicit arguments

#		i-arguments						
		0	**1**	**2**	**3**	**4**	**5**	**total**
e-arguments	**0**	2,296	3,251	3,050	412	8	1	9,018➔ 49.57%
	1	3,127	4,009	720	53	4	0	7,913➔43.5%
	2	925	262	44	2	0	0	1,233➔6.78%
	3	25	3	0	0	0	0	28➔0.15%
	4	0	0	0	0	0	0	0➔0%
	5	0	0	0	0	0	0	0➔0%
	total	6,373 ↓ 35.03%	7,525 ↓ 41.36%	3,814 ↓ 20.97%	467 ↓ 2.57%	12 ↓ 0.07%	1 ↓ 0.01%	18,192➔100%

Table 8 presents the frequencies of explicit and implicit arguments per argument position and thematic role. As shown in Table 8, arg0 is the most frequently realized implicit argument and corresponds to the agent role in 91.24% of cases (7,511 instances), followed at some distance by the cause role (800 instances, 8.70%). However, arg1 is realized slightly more frequently as an explicit argument than as an implicit argument (6,428 arg1 vs. 6,189 iarg1 instances respectively). In both cases, this argument corresponds mainly to the patient role (4,102 instances 63,81% of arg1 and 3,856 instances 62,30% of iarg1), followed by the theme role (2,274 instances 35.37% of arg1 and 2,301 instances 37.16% of iarg1). It is worth mentioning that arg2 is the third most frequently realized implicit argument, especially in a beneficiary role (800 instances, approximately 42% of cases).

Table 8 Instances of explicit and implicit arguments per thematic roles

	e-arguments						i-arguments				
	arg0	**arg1**	**arg2**	**arg3**	**arg4**		**iarg0**	**iarg1**	**iarg2**	**iarg3**	**iarg4**
agt	2,874	-	-	-	-	**agt**	7,511	-	-	-	-
atr	-	-	112	-	-	**atr**	-	-	93	-	-
ben	-	-	149	3	-	**ben**	-	-	800	22	-
cau	103	-	-	-	-	**cau**	717	-	-	-	-
cot	-	-	63	-	-	**cot**	2	-	130	-	-
des	-	-	-	-	97	**des**	-	-	-	1	99
efi	-	-	45	3	17	**efi**	-	-	40	0	67
ein	-	-	50	1	-	**ein**	-	-	73	17	-
exp	4	-	12	-	-	**exp**	1	-	38	-	-
ext	-	1	89	-	-	**ext**	-	-	84	-	-
ins	-	-	4	1	-	**ins**	-	-	7	1	-
loc	-	51	236	-	-	**loc**	-	32	275	14	-
ori	-	-	-	33	-	**ori**	-	-	-	73	-
pat	-	4,102	-	-	-	**pat**	-	3,856	-	-	-
src	8	-	-	-	-	**src**	1	-	-	-	-
tem	-	2,274	284	-	-	**tem**	-	2,301	368	-	-

2. Most implicit arguments are located near their referenced discourse entity. As shown in the last row of Table 9, 68.15% of the total number of implicit arguments annotated are located within the sentence containing the nominal predicate, 23.41% are found within the previous sentence or sentences (12.04% and 11.37% respectively), and 8.44% in the sentence or sentences (4.55% and 3.89% respectively) following the deverbal noun. This can be explained in terms of discourse coherence. That is, in order to avoid redundancy, when participants appear in the same or in the surrounding sentences, the argument is implicit because it can be easily inferred from the nearby context.

   Table 9 contains the percentages of implicit arguments (column 1) realized in the same sentence (= 0, column 2), in the previous sentence (=-1, column 3), in previous sentences, i.e. in 2 or more previous sentences (<-1, column 4), in the subsequent sentence (=+1, column 5), and in more than 2 subsequent sentences (>+1, column 6). The last column includes the total number of instances per argument position. The last row of the table shows the total percentage for referenced entity location per argument position.

   Table 9 Percentages of the distance between the referenced discourse entity and the implicit argument position

i-args	=0	=-1	<-1	=+1	>+1	total
iarg0	6,025➔69.82% ↓ 51.06%	1,016➔11.77% ↓ 48.75%	871➔10.09% ↓ 44.24%	398➔4.61% ↓ 50.57%	319➔3.7% ↓ 47.33%	8,629➔100%
iarg1	4,417➔68.70% ↓ 37.44%	744➔11.57% ↓ 35.70%	777➔12.09% ↓ 39.46%	257➔4% ↓ 32.66%	234➔3.64% ↓ 34.72%	6,429➔100%
iarg2	1,211➔61.88% ↓ 10.2%	271➔13.85% ↓ 13%	265➔13.54% ↓ 13.46%	115➔5.88% ↓ 14.61%	95➔4.85% ↓ 14.09%	1,957➔100%
iarg3	69➔53.49% ↓ 0.58%	22➔17.05% ↓ 1.06%	25➔19.38% ↓ 1.27%	6➔4.65% ↓ 0.76%	7➔5.43% ↓ 1.04%	129➔100%
iarg4	77➔45.56% ↓ 0.65%	31➔18.34% ↓ 1.49%	31➔18.34% ↓ 1.57%	11➔6.51% ↓ 1.4%	19➔11.24% ↓ 2.82%	169➔100%
Total	11,799➔68.15% ↓ 100%	2,084➔12.04% ↓ 100%	1,969➔11.37% ↓ 100%	787➔4.55% ↓ 100%	674➔3.89% ↓ 100%	17,313➔100%

3. Most of the implicit arguments can be recovered from coreference chains. We have observed that 76.69% of all implicit arguments are retrieved from previously annotated entities in coreference chains, as shown in the last row of Table 10. The remaining 23.30% correspond to entities that appear just once in the document (singletons). These data are interesting especially for implicit semantic role labeling systems, which usually use coreference information as a feature to detect implicit arguments. It is worth noting that 23.30% of implicit arguments, which is not a negligible figure, could not be resolved if we only take into account the coreference chains previously annotated, leaving aside the singletons.

4. Finally, deverbal noun arguments are more optional than verbal arguments, in the sense that verbal arguments are more explicitly realized than nominal arguments. Table 11 contains a comparison of the percentages of arguments (including explicit and implicit[36] arguments) realized in verbal predicates and in nominal predicates (second and third columns). The figures show that 23.29% of verbs appear without any explicit argument whereas 12.62% of nouns do not realize an explicit or implicit argument. Verbs with one argument explicitly realized represent 26.33%, whereas nouns represent 35.06% including both explicit and implicit arguments. Another interesting fact is that almost 46% of verbs have two explicit arguments, and almost 44% of nouns also have two arguments.

[36] It is worth noting that the implicit arguments of verbs are not annotated in Iarg-AnCora, so the number of occurrences and percentages for verbs only includes explicit arguments.

Table 10 Percentages of implicit arguments realized as mentions in a coreference chain (entity label) or singletons

i-args	entity	singleton	total
iarg0	7,092➔82.09% 🡻 53.34%	1,547➔17.91% 🡻 38.29%	8,639➔100%
iarg1	4,669➔72.51% 🡻 35.12%	1,770➔27.49% 🡻 43.81%	6,439➔100%
iarg2	1,366➔69.73% 🡻 10.27%	593➔30.27% 🡻 14.68%	1,959➔100%
iarg3	75➔58.14% 🡻 0.56%	54➔41.86% 🡻 1.34%	129➔100%
iarg4	93➔55.03% 🡻 0.7%	76➔44.97% 🡻 1.88%	169➔100%
Total	13,295➔76.69% 🡻 100%	4,040➔23.30% 🡻 100%	17,335➔100%

Table 11 Percentages of (explicit and implicit) arguments per verbs and nouns

args	(e-args) verbs	(e-args+i-args) nouns
0	12,970 (23.29%)	2,296 (12.62%)
1	14,660 (26.33%)	6,378 (35.06%)
2	25,576 (45.93%)	7,984 (43.89%)
3	2,458 (4.41%)	1,419 (7.8%)
4	21 (0.04%)	108 (0.59%)
5	1 (0%)	7 (0.04%)
total	55,686 (100%)	18,192 (100%)

6 Applications: Applying Iarg-AnCora for building LIARc

A subset of Iarg-AnCora has already been used as a training and test corpus for creating LIARc (Peris et al. 2013), a supervised ML feature-based model for detecting the core implicit arguments of deverbal nominalizations based on linguistically informed features. It was the first system to deal with this type of arguments in Spanish. LIARc basically replicated the experiments carried out for English by Gerber and Chai (2010, 2012) and proposed a number of variations and improvements on the features used. We selected the G&C model because it was easier to scale up and did not suffer from the data-sparseness problem found in the SemEval training corpus. The eight most frequent unambiguous deverbal nominalization lemmas in Iarg-AnCora were selected for the building of LIARc. Unambiguous lemmas were those deverbal nouns that only have one sense and one associated syntactic-semantic frame. We selected eight predicates and not ten like in G&C model because there is a severe drop in frequency from the ninth predicate in Iarg-AnCora. This set of eight nominalizations corresponds to a total of 469 instances shown in the first two columns of Table 12. The remaining four columns present the total number and average number of implicit arguments for each predicate and the same information for explicit arguments.

The main problem for implicit arguments detection approaches is the lack of appropriate learning corpora and the sparseness of the existing ones (such as SemEval). G&C tackle this problem by focusing on a reduced set of the 10 most frequent monosemous deverbal nominalizations in English for which a quite dense dataset has been built (see Table 1). G&C's method is based on using supervised MLs for training their own corpus. Some of the features included are widely used in most of the systems, including the distance between the predicate and the candidate implicit argument, the local context of both, statistical coocurrence measures (such as PMI), and properties of the verb from which the nominalization derives,

among others. Besides these features, G&C uses highly lexicalized ones, such as the word forms and lemmas of the nominalization and the argument, and their local context, as well as generalizations using WordNet.

The number of explicit and implicit core arguments is shown for each predicate. These figures are much higher than those reported by G&C for English. Although the figures are not directly comparable, the difference is due to both the lower number of explicit arguments for Spanish nominalizations (0.5 on average)[37] compared to English (1.1 in G&C) and to the higher number of implicit arguments (1.3 vs. 0.8). The average number of implicit arguments per predicate instance is 1.3 with a standard deviation of 0.41 (0.8 in G&C)[38].

Table 12 Statistics for the eight deverbal nouns in Iarg-AnCora (Peris et al. 2013)

Deverbal noun	**Instances**	**Nbr. i-args**	**Avg. i-args**	**Nbr. e-args**	**Avg. e-args**
actuación, actuation	67	35	0.5	29	0.4
comunicado, communication	63	116	1.8	16	0.2
daño, harm	42	67	1.6	12	0.3
empate, draw	41	54	1.3	12	0.2
negocio, business	50	59	1.2	16	0.3
opción, option	48	56	1.2	34	0.7
propuesta, proposal	104	177	1.7	78	0.8
viaje, travel	54	66	1.2	30	0.5
TOTAL	**469**	**630**	**1.3**	**227**	**0.5**

In AnCora-Es, for each document *d*, the implicit arguments of a nominalization instance are annotated at entity level, i.e. given a predicate instance *p*, and an implicit argument tag *iargn="r"*, the *iargn="r"* filler, if existing, is annotated with the identifier of an entity $e \in E$, E being the set of entities occurring in *d*. Entities in *E* can be regular ones, occurring in coreference chains, or those appearing just once (singletons). We have observed that 75% of core implicit arguments are retrieved from regular entities while the remaining 25% correspond to singletons.

For a regular entity *e*, the set of mentions (the coreference chain) is noted as $e' = \{e_1, \ldots, e_i, \ldots, e_n\}$. Usually, at least one mention e_i, in the coreference chain e', is an explicit argument of a predicate. When *e* is a regular entity, mentions in e' tend to be NPs, while in the case of singletons other possibilities exist (prepositional, adjectival, adverbial phrases, possessives or subordinate clauses).

Our learning setting tries to tackle two challenges that are difficult to make compatible: 1) to follow G&C's proposal as closely as possible given the highly accurate performance of its features and the possibility of comparing their results with ours; and 2) to learn LIARc for the semi-automatic annotation of the whole AnCora-Es, i.e, to scale up to the whole corpus. These challenges are difficult to make compatible because most of the best features used by G&C are highly lexicalized.

Hence, the features for the LIARc classification model were inferred from this training corpus. We experimented with four models depending on whether the features were lexicalized or not, and whether the features were specific or generalized: 1) lexicalized-specific; 2) lexicalized-generalized; 3) non-lexicalized-specific, and 4) non-lexicalized-generalized, of which only the last one was used for building the LIARc classifier. Lexicalized features contained, for instance, the specific predicate involved and the words, lemmas, synsets and predicates surrounding the antecedent to be linked; whereas in the non-lexicalized models a null string replaced the lexicalized features. Specific features contain, for instance, WN synsets corresponding to the nominalization itself or its verbal origin. Using the kind of features presented by G&C involves, in practice, learning a classifier for each lemma. This schema is, obviously, impossible to scale up. Generalization was performed through synsets occurring in the features by the

[37] The figures are slightly different from those reported in section 4 because the comparison with G&C is performed with the subset of the 8 most frequent monosemous nominalizations.
[38] A third explanation could be the use of different criteria in the annotation of both explicit and implicit arguments in the G&C dataset and in AnCora.

Top Concept Ontology (TCO) labels[39] (Álvez et al. 2008) attached to them,[40] in the case of nouns, and, in the case of verbs, their lemmas were replaced by the semantic classes of AnCora-Verb.

Details on the learning process can be found in Peris et al. (2013). For learning, we used a supervised machine learning approach. Due to the highly lexicalized nature of most of the features, and their precision oriented type, we chose Adaboost as a classifier model. Due to the highly unbalanced distribution of positive and negative examples, we weighted the positive ones by a factor of 69, leading to a more balanced distribution. As in G&C, we used feature templates for generating the features.

For instance, the most accurate features were derived from the feature template <pe_i, arg_i, p, $iarg_n$>. This template has to be instantiated for all the possible values of the predicate p (8 values), for all the possible values of the $iarg_n$ tags (up to 36 argument positions + thematic roles), for all the possible values of the arg_i (up to 6 argument positions) and for all the possible values of pe_i. Obviously not all the combinations occur in the learning corpus. Other features include the distance between the predicate and the implicit argument (both in sentences and tokens), the verbal entry from which the predicate is derived, and several kinds of generalizations. See Peris et al. (2013) for details on the features and their accuracies.

Many of the features and feature templates we used for learning the LIARc were replicated from G&C. As these authors rank their features by accuracy, we took their most accurate features. In some cases, our features basically reproduce theirs, while, in others, our features are often simply inspired by theirs. We have selected the most accurate templates as the core of our system. From the 10 most relevant ones, we reproduced verbatim 5, including the 3 most accurate, while the other 4 are heavily based on the corresponding ones, with changes due to the specific differences between English and Spanish and the available resources (see Peris et al., 2013 for details).

The overall F-Measure for all the models was, on average, 89.9%, showing that there were no significant differences between lexicalized and non-lexicalized models. This could be explained by the fact that, unlike in G&C, none of the lexicalized features occurred among the top ranked ones. The results obtained were better than those reported by G&C for English. In addition to the previously mentioned differences between English and Spanish, discussed in section 5, this could be explained by the fact that the explicit arguments in G&C were automatically obtained, while they were manually annotated in Iarg-AnCora.

In its initial setting, learning was performed using cross-validation and the whole available material was therefore used for both learning and testing and the results reported in Peris et al. (2013) are based on this evaluation. Later, after the whole Iarg-AnCora corpus had been built, an additional test, using the set of iargs corresponding to lemmas not included in the set of 8 monosemous ones used for learning was performed. The results obtained outperformed the initial ones.

Once the building of Iarg-AnCora was finished, we used the corpus for training a new version of LIARc. But this time, we applied the non-lexicalized-generalized model using for the whole corpus for learning, that is, the whole set of nominalizations (1,454 different lemmas corresponding to 18,397 instances), and not only the 8 unambiguous lemmas used previously. The results were better than those obtained in the previous experiment with an F-Measure of 92.91%. Hence, this last model increased the F-Measure performance 3.1 points over the previous one confirming, on the one hand, that the process of delexicalization seems to have a very limited effect on the figures obtained, and, on the other, that the large increase in training data improves the results obtained by the systems.

[39] TCO aims to provide WordNet synsets with a neutral ontological assignment. The ontology contains 63 features organized as 1st order entities (physical things), 2nd order entities (situations) and 3rd order entities (unobservable things).

[40] Since AnCora-Es mentions are annotated with correct synsets, no Word Sense Disambiguation was needed.

7 Conclusions

In this paper, we have described the annotation of the Spanish Iarg-AnCora corpus with the implicit arguments of deverbal nouns, focusing on the methodology used and the annotation scheme adopted. In Iarg-AnCora, the core implicit arguments are linked to their corresponding discourse entities (i.e. to all mentions in a coreference chain or to a singleton), where there exists an identity relation between the antecedent (i.e. a discourse entity) and the implicit argument. The annotation scheme combines those schemes used in the annotation of verbal argument structure and coreference in the AnCora corpus, which follow in turn the PropBank argument structure scheme and the general criteria of the MATE coreference scheme. The results of the inter-annotator agreement test conducted are also presented. The agreement obtained (81% observed agreement) ensures the reliability of the annotation, and the analysis of disagreements enabled us to detect and resolve the errors. Most of the disagreements were related to missed links between the argument and the discourse entity, and to different interpretations of the antecedent selected for linking the argument. We also presented the specialized interface designed for annotating the implicit arguments integrated in AnCoraPipe. This tool allows us to tag different layers of linguistic information and the XML output files containing all the linguistic -morphological, syntactic, semantic and discourse- information.

Our main motivation for building Iarg-AnCora was to provide a corpus annotated with implicit arguments with a wide coverage in order to avoid the problem of data sparseness found in the available English corpora. Our main goal was to have a reference resource to study the implicit arguments of Spanish deverbal nouns empirically. Iarg-AnCora contains 18,397 nominal tagged instances, corresponding to 1,454 different deverbal nouns, with an average of 0.64 implicit arguments per predicate. Since we annotated an already existing corpus, the data have different ratios of instances per word token, ranking from 225 to 1 instance per lemma (where 30% of lemmas have more than 10 occurrences, 48.3% more than 5, and 25% have only 1 occurrence per lemma). Therefore, each lemma does not have a comparable density of annotation. The annotation of these arguments results in an important gain in role coverage (128% on average in the annotation of explicit arguments). Therefore, if we do not take into account the implicit arguments, relevant semantic information is missed, and this missing information is crucial for a better understanding of sentences and, consequently, for a better understanding of texts. The analysis of the annotated corpus confirms our initial hypotheses: a) implicit arguments are more frequent than explicit arguments in deverbal nominalizations, with the most common being the arguments closest to the predicate (i.e. arg0 and arg1); b) most implicit arguments are located near their referenced discourse entity; they usually appear within the sentence containing the nominal predicate; c) most implicit arguments can be recovered from coreference chains and, d) verbal arguments are more often explicitly realized than deverbal noun arguments, almost 50% of which do not express an argument explicitly.

All this tagged information can also be very useful for training, developing and testing SRL and CR systems, which can use different learning features. For instance, we used Iarg-AnCora to train the LIARc classifier, which is based on linguistic features obtained from the corpus, and the results obtained indicate that the performance of the classifier improves when the training data is increased and the sparseness of the data is reduced. Our next aim is to develop a SRL system dealing with both nominal and verbal predicates, which will take into account the discourse context. This SRL system could be learned from Iarg-AnCora. Another line of work would be to assess the possibility of applying LIARc to AnCora-Cat, the Catalan version of AnCora-Es as an initial step for building automatically Iarg-AnCora-Cat. It would also be interesting to enrich Iarg-AnCora with the annotation of the implicit arguments of verbs, and also to tag the non-resolvable arguments.

Acknowledgments

We are grateful 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 DIANA (TIN2012-38603-C02-02) and SKATER (TIN2012-38584-C06-01) projects from the Spanish Ministry of Economy and Competitiveness.

References

Álvez, J., Atserias, J., Carrera, J., Climent, S., Oliver, A., & Rigau, G. (2008). Consistent annotation of EuroWordnet with the top concept ontology. *Proceedings of Fourth International WordNet Conference (GWC -08),* Association for Computational Linguistics.

Aparicio, J., Taulé, M., & Martí, M.A. (2008). AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. *Proceedings of 6th International Conference on of Language, Resources and Evaluation*, Marrakech, Morocco.

Baker, C.F., Fillmore, C.J., & Lowe, J.B. (1998). The Berkeley FrameNet Project. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics,* Vol. 1. ACL'98, Stroudsburg, PA, USA, Association for Computational Linguistics, 86–90.

Balvet, A., Condette, M.H., Haas, P., Huyghe, Marín R., & Merlo, A. (2011). Nomage: an electronic lexicon of French deverbal nouns based on a semantically annotated corpus. *Proceedings of the First International Workshop on Lexical Recources (WoLeR 2011),* 8-15.

Bertran, M., Borrega, O., Martí, M.A, & Taulé, M. (2011). AnCoraPipe: A new tool for corpora annotation. Working paper 1: TEXT-MESS 2.0 (Text-Knowledge 2.0). Universitat de Barcelona. http://clic.ub.edu/sites/default/files/pagines/AnCoraPipe.pdf

Chen, D., Schneider, N., Das, D., & Smith, N.A. (2010). SEMAFOR: Frame argument resolution with log-linear models. *Proceedings of the 5th International Workshop on Semantic Evaluation* (*SemEval'10),* Stroudsburg, PA, USA, Association for Computational Linguistics, 264–267.

Chinchor, N., & and Sundheim, B. (2003). *Message Understanding Conference (MUC) 6*. Linguistic Data Consortium, Philadelphia.

Erk, K., & Padó, S. (2004). A powerful and versatile XML Format for representing role-semantic annotation. *Proceedings of 4th International Conference on Language Resources and Evaluation.* Lisbon, Portugal.

Fillmore, C.J. (1986). *Pragmatically Controlled Zero Anaphora.* Technical report, Department of Linguistics. University of California.

Fillmore, C.J., & Baker, C.F. (2001). Frame semantics for text understanding. *Proceedings of the Workshop on WordNet and Other Lexical Resources*, NAACL, Pittsburgh, Pennsylvania, Association for Computational Linguistics.

Gerber, M., &. Chai, J.Y. (2010). Beyond NomBank: a study of implicit arguments for nominal predicates. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL'10*, Stroudsburg, PA, USA, Association for Computational Linguistics, 1583–1592.

Gerber, M. (2011). Semantic role labeling of implicit arguments for nominal predicates. Ph-Dissertation, Michigan State University, USA.

Gerber, M., & Chai, J.Y. (2012). Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Computational Linguistics,* 38, 755–798.

González-Agirre A., Laparra E., & Rigau G. (2012). *Multilingual Central Repository version 3.0*. *Proceeding of 8th international conference on Language Resources and Evaluation (LREC'12)*. Istambul, Turkey.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: The 90% Solution. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL'06*, 57–60, New York.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2006). Extending VerbNet with novel verb classes. *Proceedings of the 5th international conference on language resources and evaluation (LREC'06),* 1027–1032, Genova, Italy.

Laparra, E., & Rigau, G. (2012). Exploiting Explicit Annotations and Semantic Types for Implicit Argument Resolution. *ICSC*, 75–78.

Laparra E., & Rigau G. (2013). ImpAr: A Deterministic Algorithm for Implicit Semantic Role Labelling. *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Sofia, Bulgaria. Aquest crec que es unsupervised.

Levin, B. (1993). *English Verb Classes and Alternations: A preliminary investigation*. University of Chicago Press, Chicago.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19, 313-330.

Meyers, A., Reeves, R., & Macleod, C. (2004). NP-external arguments, a study of argument sharing in English. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE'04)*, Stroudsburg, PA, USA. Association for Computational Linguistics, 96–103.

Meyers, A. (2007). Anotation Guidelines for NomBank-Noun Argument Structure for PropBank. Technical report, University of New York.

Mitchell, A., Strassel, S., Przybocki, M., Davis, JK, Doddington, G., Grishman, R., Meyers, A., Brunstein, A., Ferro, L., & Sundheim, B. (2003). *ACE-2 Version 1.0*. Linguistic Data Consortium, Philadelphia.

Moor, T., Roth, M., & Frank, A. (2013). Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. *Proceedings of the 10th International Conference on Computational Semantics (IWCS) - Short Papers*, Potsdam, Germany, 369-375.

Müller, H. (2011). The Copenhagen Dependency Treebank (CDT). Extending syntactic annotation to morphology and semantics. In K. Gerdes, E. Hajičová, L. Wanner (Ed.), *Depling 2011 Proceedings. International Conference on Dependency Linguistics: Exploring Dependency Grammar, Semantics, and the Lexicon*,125-134. Barcelona: Depling.

Palmer, M.S., Dahl, D.A., Schiffman, R.J., Hirschman, L., Linebarger, M., & Dowding, J. (1986). Recovering Implicit information. *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, New York, USA, Association for Computational Linguistics, 10–19.

Palmer, M., Kingsbury, P., & Gildea, D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, 21 (1).

Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2011). *English Gigaword Fifth Edition*. Linguistic Data Consortium, Philadelphia.

Peris, A., & Taulé, M. (2011). AnCora-Nom: A Spanish Lexicon of Deverbal Nominalizations. Procesamiento del Lenguaje Natural, 46: 11–19.

Peris, A., & Taulé, M. (2012). Annotating the Argument Structure of Deverbal Nominalizations in Spanish. *Language Resources and Evaluation*, 46 (4): 667-699, Springer-Verlag.

Peris, A., & Taulé, M. (2013). *Argumentos implícitos de los sustantivos deverbales. Guía de anotación v. 0.2*. Working Paper: 1 Diana-Construcciones. Universitat de Barcelona.

Peris, A., Taulé, M., Rodríguez, H., & Bertran, M. (2013). LIARc: Labeling Implicit ARguments in Spanish deverbal nominalizations. *Computational Linguistics and Intelligent Text Processing - 14th International Conference*, *CICLing 2013*, Samos, Greece. Proceedings, Part I. Springer, *Lecture Notes in Computer Science*, 7816: 423-434, Berlin, Germany.

Poesio, M. (2004). The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial workshop at HLT-NAACL 2004*: 154–162. Boston.

Poesio, M., Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 76–83, Ann Arbor, MI.

Recasens, M., & Martí, M. A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation,* 44 (4): 315-345, Springer-Verlag.

Recasens, M., & Vila, M. (2010). On Paraphrase and Coreference. *Computational Linguistics,* 36(4): 639-647.

Roth, M., & Frank, A. (2012). Aligning Predicate Argument Structures in Monolingual Comparable Texts: A New Corpus for a New Task. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada, 218–227, Association for Computational Linguistics.

Roth, M., & Frank, A. (2013). Automatically identifying implicit arguments to improve argument linking and coherence modeling. *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*, Atlanta, Georgia, USA, 306-316, Association for Computational Linguistics.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C.R., & Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. Berkeley, California: International Computer Science Institute.

Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., & Palmer, M. (2010). Semeval-2010 task 10: Linking events and their participants in discourse. *Proceedings of the 5th Workshop on Semantic Evaluations (ACL 2010)*, 45-50, Uppsala, Sweden.

Ruppenhofer, J., Gorinski, P., & Sporleder, C. (2011). In search of missing arguments: A linguistic approach. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2011)*: 331–338*,* Hissar, Bulgaria.

Ruppenhofer, J., Lee-Goldman, R., Sporleder, C., & Morante, R. (2012). Beyond sentence-level semantic role labeling: linking argument structures in discourse. *Language Resources and Evaluation*, 47 (3): 695-721, Springer-Verlag.

Silberer, C., & Frank, A. (2012). Casting Implicit Role Linking as an Anaphora Resolution Task. **SEM 2012: The First Joint Conference on Lexical and Computational Semantics* – Vol.1: *Proceedings of the main conference and the shared task*, and Vol. 2: *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012),* Montréal, Canada, Association for Computational Linguistics, 1–10.

Taulé, M., Martí, M.A., & Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. *Proceedings of 6th International Conference on Language Resources and Evaluation,* 96-101, Marrakesh, Morocco.

Taulé, M., Martí, M.A., & Borrega, O. (2011). AnCora 2.0: Argument Structure Guidelines for Catalan and Spanish, Working paper 4: TEXT-MESS 2.0 (Text-Knowledge 2.0).

Tetreaul, J.R. (2002). Implicit Role Reference. *International Symposium on Reference Resolution for Natural Language Processing*, 109–115.

Tonelli, S., & Delmonte, R. (2010). VENSES++: Adapting a deep semantic processing system to the identification of null instantiations. *Proceedings of the 5th Inter- national Workshop on Semantic Evaluation* (*SemEval'10)*, Stroudsburg, PA, USA, Association for Computational Linguistics, 296–299.

Tonelli, S., & Delmonte, R. (2011). Desperately seeking implicit arguments in text. *Pro- ceedings of the ACL 2011 Workshop on Relational Models of Semantics*, Stroudsburg, PA, USA, Association for Computational Linguistics, 54–62.

Wang, N., Li, R., Lei, Z., Wang, Z., & Jin, J. (2013). Document Oriented Gap Filling of Definite Null Instantiation in FrameNet, M. Sun et al. (Eds.), *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data 2013, Lecture Notes in Computer Science,* 85–96, Springer-Verlag Berlin Heidelberg.

Weischedel, R., Hovy, E., Marcus, M., Palmer M., Belvin, R., Pradhan, S., Ramshaw, L., & Xue, N. (2011). OntoNotes: A Large Training Corpus for Enhanced Processing. In J. Olive, C. Christianson & J. McCary (Eds.), *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.