# Definite Description Resolution in Spanish

Rafael Muñoz and Antonio Ferrández
Grupo de Procesamiento del Lenguaje y Sistemas de Información
Universidad de Alicante
Campus de San Vicente del Raspeig, Alicante, Spain
{rafael, antonio}@dlsi.ua.es

## Abstract

*In this work, a method for the resolution of the identity co-reference produced by definite descriptions in Spanish texts is presented. This method is based on the linguistic knowledge acquired by POS-tagger, synonymous dictionary and relationships between names and verbs resources. This method uses a system of restrictions and preferences in order to obtain the correct antecedent. The method achieves a precision of 95% for the identification of non-anaphoric and anaphoric uses of definite descriptions. Moreover, the method achieves a precision of 80% for the reference resolution task. In addition, a profound study of linguistic characteristics of definite descriptions in a Spanish corpus has been made. From this study, a classification of the different definite description types has been obtained.*

## 1. Introduction and related works

Anaphora resolution is one of the most important and difficult problems to be solved in any natural language processing system. There are several grammatical categories with referential properties such as pronouns, adverbs, adverbial adjuncts and definite descriptions. This paper focuses on resolving definite descriptions. Definite description is defined as a noun phrase headed by the definite article, such as *the dog*.

Processing of definite descriptions has been pointed out by several authors. Among them, two methods can be emphasised. Firstly, a pronominal anaphora and definite description processing algorithm is described by Kameyama in [6]. This algorithm is applied to the information extraction system FASTUS [5]. This information extraction system works in a general geo-political business domain, in which its anaphora resolution algorithm only solves *identity co-reference* without evaluating other co-reference types as *parts of, set - subset and set - member*. Identity co-reference occurs when an anaphoric expression can be replaced by antecedents without changing the sense of the sentence.

Secondly, a system that resolves definite description is presented by Vieira and Poesio in [14] and [13]. This system solves references between definite descriptions and antecedents with either the same head noun or a semantic relationship (synonym, hyperonym, hyponym). Clark, in [2], named *bridging references* the uses of definite descriptions whose antecedents have a different head noun.

Previous methods focused on solving definite descriptions in the English language. Any method for the resolution of the definite descriptions in unrestricted texts in the Spanish language does not exist. The only method that exists works in a restricted domain. This method developed by Muñoz et al. in [9] is applied to the information extraction system EXIT [7].

This paper is organized as follows. In section two, a classification of definite descriptions that are used in written texts is presented. This classification is based on their referential or non-referential characteristics. Referential definite descriptions are also classified from the kind of knowledge that is used to establish the relationship between anaphoric expressions and antecedents. In section three, an algorithm to solve definite descriptions in unrestricted Spanish text is proposed. This algorithm deals with identity co-references between antecedent and anaphoric expression (definite description). In section four, a system's evaluation from a fragment of LEXESP corpus is made.

## 2. Classification of definite descriptions

Several taxonomies as Cristopherson [1], Hawkins [4], Prince [12] and Poesio & Vieira [11] have been carried out for definite descriptions in the English language. These taxonomies are usually based on relations between the anaphoric expression (definite descriptions) and the antecedent. Moreover, Poesio and Vieira's classification distinguishes an additional category using information about knowledge that speaker and listener have about the object that they are talking about. In the Spanish language, no definite description taxonomy has been accomplished. One of the main differences between English and Spanish definite descriptions is the usual omission of the determiner *the* in the English language, whereas in Spanish it is usual to add it to all noun phrases (determiner *el/la*). For example, the Spanish sentence *La generosidad es uno de los valores ms importantes* is translated into English as $\emptyset_{the}$ *Charity is one of the most important values*, where $\emptyset_{the}$ symbol points the position of the determiner the. This fact is due to the fact that definite descriptions in English are only used to refer to a previously mentioned entity. Therefore, the number of Spanish definite descriptions is greater than English ones.

In the taxonomy presented by Poesio and Vieira in [11], all definite descriptions can be classified, although they cannot be classified from the kind of knowledge that is needed to solve the reference between anaphoric expression and its antecedent.

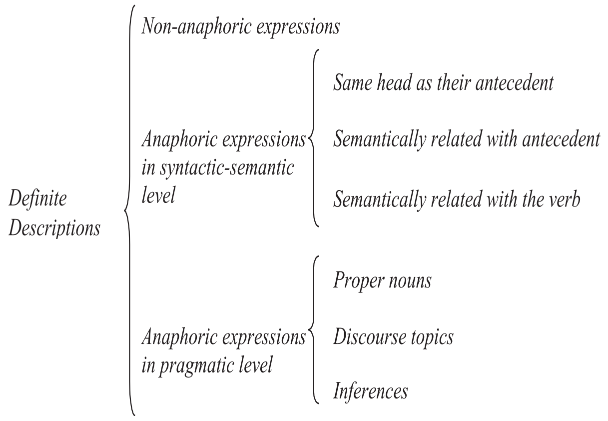


**Figure 1. Definite description's classification**

In Figure 1 a definite description (*DD*) classification obtained from the LEXESP corpus is shown. This study has been accomplished according to the linguistic level of the kind of knowledge needed to solve the anaphoric expression. There are four linguistic levels: phonologic, morphologic, syntactic-semantic-textual and pragmatic level. LEXESP is a Spanish corpus formed by a set of narrative texts written by different authors. This corpus has about 5 million of tagged words developed by Psychology Department from University of Oviedo, Computational Linguistic Group of University of Barcelona and Language Treatment Group from Technical University of Catalua.

Next, each category in Figure 1 is going to be described in full detail:

**a) Non-anaphoric expressions.** Within this category, those definite noun phrases that have no reference meaning with other previously mentioned entities are included. These noun phrases usually introduce a new entity that can be referenced by other noun phrases. For example, let us suppose a text that starts with the sentence *La casa de la playa es bonita (The house on the beach is beautiful)*. In this case, the definite description *The house on the beach* has not previously appeared in discourse, therefore, it introduces a new entity.

**b) Anaphoric expressions in syntactic-semantic level.** In this category, those definite noun phrases that reference other previously mentioned entities are included. According to the kind of syntactic-semantic relation that is established between anaphoric expression and its antecedent, we can distinguish between the following types:

1. **Anaphoric expressions with the same head as their antecedent.** That is to say, those noun phrases that share their head noun. For example, in the following text: *La casa de la playa es bonita ... La casa era de un mdico (The house on the beach is beautiful ... It was the house of a doctor)*, the underlined noun phrases share the same head noun, i.e. *casa (house)*. Those Spanish noun phrases with their head omitted[1] are also included in this category. For example, *La del mdico tiene piscina (The doctor's one has a swimming-pool)*.

2. **Anaphoric expressions semantically related with their antecedent.** It occurs when the head noun of the definite description is a *synonym*, *hyponym* or *meronym* of its antecedent's head. For example, in *Los soldados irrumpieron en el poblado. Estos hombres valientes... (The soldiers burst into the village. These brave men...)*, the definite description *estos hombres valientes (these brave men)* refers to *los soldados (the soldiers)* since *soldiers* is an hyponym of *men*.

3. **Anaphoric expressions semantically related with the verb.** This kind of expressions refers to an antecedent that appears in a sentence whose verb means the action developed by the description. For example, in *Juan vende su casa a Fernando. El vendedor pagar las costas (Juan sells his house to Fernando. The seller will pay the expenses)*, the definite description *el vendedor (the seller)* refers to *Juan* since it is the agent of the action of the verb *to sell*.

**c) Anaphoric expressions in pragmatic level.** This category includes definite descriptions that indirectly refer

[1]This kind of noun phrases are quite usual in Spanish texts. Its equivalent in English would be one-anaphora.

to their antecedents. In order to solve these references the knowledge of the world is needed. We can distinguish between the following kinds of references:

1. **Proper nouns**, i.e. definite descriptions whose antecedent is a proper noun. Moreover, they usually mean some characteristics of their antecedent. For example: *Bill Clinton and The President of US*.
2. **Discourse topics**, i.e. definite descriptions that refer to the discourse topic of the text in which anaphora occurs. For example, in a sports text, we can usually make references to sports-teams.
3. **Definite descriptions that can be inferred**, i.e. when the anaphoric expression has a complex relationship with its antecedent (cause-effect relation, cause-consequence relation, etc.). For example: *Los ataques eran bastante usuales. Las vctimas estaban por todas partes. (The attacks were quite usual. The victims were everywhere)*, where the victims are a consequence of the attacks.

Syntactic-semantic-textual levels include definite descriptions named by Poesio and Vieira as the *same head anaphora* and some of their *associative uses* (*synonym, hyponym, meronym* and *events*) in [11]. The world knowledge is not needed in order to solve *anaphoric expressions in syntactic-semantic level*. It can be solved by means of syntactic knowledge and semantic knowledge (e.g. knowledge obtained from WordNet or EuroWordNet that provide us with information about *synonym, hyponym* or *meronym*).

In Table 1, the distributions of definite descriptions in a fragment of LEXESP corpus are shown. It should be remarked that 52% of definite descriptions do not refer to any previous noun phrase, i.e. they introduce a new entity in discourse. The 18% of definite descriptions need world knowledge to find out their antecedent (pragmatic level) and the remaining 30% use syntactic and semantic knowledge (synonym, relation between verb and noun, etc.).

| **Expression** | **Percentage** |
|---|---|
| Non -anaphoric expressions | 52% |
| Anaph. expr. in syntactic-semantic level | 30% |
| Anaph. expr. in pragmatic level | 18% |

**Table 1. Distribution of definite descriptions in LEXESP corpus**

Among anaphoric expressions in syntactic-semantic-textual level, such Table 2 shows, the 81% of definite descriptions have the same head-noun as their antecedents. The 17% has a synonymy-hyperonymy relation between both heads, and for the remaining 2% there is a relation between antecedent's verb and definite description.

| **Kind of definite description** | **Percentage** |
|---|---|
| Same Head | 81% |
| Semantic relation | 17% |
| Verb relation | 2% |

**Table 2. Distribution of anaphoric expressions in syntactic-semantic-textual level**

## 3. Algorithm to solve definite descriptions

This algorithm takes as input the slot structure (*SS*) of the sentence and a list of antecedents (*LA*) that have appeared in previous sentences. It returns a list of antecedents (*LASal*) as output in which the definite description is added if it introduces a new entity, or otherwise its antecedent is added. In Figure 2, an algorithm to solve definite descriptions is shown.

```
Solving DD
        use SS, LA
        return LASal
while not end of SS
   Look for DD into SS
   If NP is a DD then
     Applying restrictions to Antecedents List
     Appllying preferences to Antecedenst List after restrictions
     If not empty Antecedents Lists after preferences then
       Adding Antecedent to Antecedents List LA, LASal)
     else
        Adding DD to Antecedents Lists
   else
      Adding NP to Antecedents Lists
End while
```

**Figure 2. Algorithm to solve definite descriptions**

Next, each step of the algorithm is going to be described in full detail. The steps of the algorithm are the following:

1. **The detection of the definite description in the text**. The algorithm takes the output of the partial parser *SUPP* as input, presented by Martínez-Barco et al. in [8] and Palomar et al. in [10], and a list of antecedents (*LA*) that appear in previous sentences. This parser returns a *slot structure (SS)* for each sentence. This *SS* stores all syntactic, lexical, morphologic and semantic knowledge for every component of sentence or subconstituent of the sentece, [3].

This algorithm considers definite descriptions as noun phrases that begin with a definite or demonstrative article. Otherwise, if the algorithm detects others noun phrase types are introduced in the list of antecedents they are considered as new entities.

The algorithm for the resolution of definite descriptions starts searching these definite descriptions in sentences. The algorithm detects definite descriptions by means of scanning the sentence's *SS* for noun phrases, i.e. arguments with name: *noun phrase*. When a noun phrase is found, then its *SS* is also scanned for a definite or demonstrative article. If the noun phrase is a definite description then the algorithm continues in the next step.

2. **The application of restrictions on the list of antecedents**. Once the anaphoric expression is detected, then a set of restrictions is applied to the list of antecedent *(LA)* in order to select antecedents that are compatible with the definite description.

   The list of antecedent is formed by the *SS* for each noun phrase that has appeared in previous sentences. This algorithm only considers noun phrases as possible antecedent and does not consider others kinds as sentences or full texts.

   In order to apply restrictions, a dictionary of synonyms and relations between verbs and names is used. This dictionary has been made by hand. The following restrictions are applied:

   - Antecedents with the same head noun,
   - Antecedents with a synonym, hyperonym or hyponym relation,
   - Antecedents that appear in a sentence with a verb that is related with the anaphoric expression.

   If these is more than one antecedent that satisfies restrictions, then preferences are applied. Otherwise, if there is an only antecedent, then it is considered as the solution of the definite description. This antecedent is added to the list of antecedents*LA* generating the output list of antecedents. Finally, if there is no antecedent that satisfies restrictions, then the definite description is supposed to introduce a new entity in discourse and it is added to the *LA* generating the output list of antecedents.

3. **The application of preferences on the list of remaining antecedents after restrictions**.

   In order to choose the most compatible antecedent with the anaphoric expression (definite description) a set of preferences is applied. This set of preferences is only applied when there is more than one antecedent satisfies the previous restrictions. These preferences are:

   - Antecedents with number and genre concordance are preferred.
   - Repetition, i.e. antecedents with the same pre and post-modifiers are preferred.
   - Antecedents with semantic relation between modifiers of antecedents and modifiers of definite description.
   - The algorithm prefers those antecedents that are related with the verb of the sentence in which the definite description appears. If this relationship is direct, then antecedents with a subject syntactic function are preferred. Otherwise, if this relationship is indirect, then those ones with a complement function are preferred.
   - Indefinite noun phrases and proper nouns are preferred to definite noun phrases.
   - Antecedents that have appeared most in the text are preferred.
   - Nearest antecedent is preferred.

   The scheme to apply preferences is the following. The first preference is applied to the list of antecedents that satisfies restrictions. If there is more than one antecedent that satisfies this preference, then the second preference is applied to these antecedents, i.e. the one that satisfies the first preference. And so on with the remaining preferences. This process is stopped when there is only one antecedent that satisfies a preference. In this case, this antecedent is considered as the solution of the definite description. Those preferences that are not satisfied by any antecedent are ignored, and the process continues with the following preference.

4. **The search of the following definite description**. The algorithm return to run the step one processing the same slot structure. If the algorithm uses up the slot structure then it takes the *SS* of the next sentence.

## 4. Evaluation

A fragment of LEXESP corpus has been chosen as test corpus. This algorithm consists of two tasks. The first one, classifies the definite description as anaphoric or non-anaphoric. The second one, looks for the correct antecedent. The algorithm has obtained a 95% precision[2] in classification task and an 80% precision in co-reference tasks. The

[2]Precision is the quotient between number of correctly solved and number of processed definite descriptions.

kind of definite descriptions that are included in this evaluation are identity co-references produced by noun phrases in categories: *Anaphoric expressions in syntactic-semantic-textual level* (section 2).

Although other grammar categories such as the pronoun use the gender and number restriction in order to eliminate a great amount of possible antecedents. We can make the conclusion that this restriction used in definite description algorithm does not eliminate great numbers of them and for this reason, it is better to use it as a preference.

The main contributions of this algorithm are as follows:

We deal with co-references produced by definite descriptions that are related with a verb. These kind of anaphoric expressions are not solved by other algorithms as developed by Vieira & Poesio and Kameyama in [14] and [6], respectively.

In the first step, all definite descriptions are considered as anaphoric expressions, i.e. the algorithm tries to find out its antecedent. If the algorithm cannot find a suitable solution, then this definite description is considered as an expression that introduces a new entity. Therefore, classification mistakes do not affect the remaining anaphoric expressions as it occurs in algorithms developed by Vieria and Poesio in [14] and [13].

Antecedents of a definite description can be classified into three types: with the same head, related with a verb and with a synonym, hyperonym or hyponym relationships. We consider the same importance to each kind of antecedent. In opposition, the algorithm developed by Vieira and Poesio searches antecedents with the same head. If antecedents with the same head are not found, then the algorithm will search antecedents with a synonym, hyperonym or hyponym relation.

In works by Vieira and Poesio, [14] and [13], a 83% precision is obtained in direct anaphora (with the same head). A 37% precision is obtained in associative anaphora (semantic relation). As Table 3 shows, our algorithm obtains better results[3] : an 85.7% in anaphora with the same head and a 75% in semantic relations. Moreover, definite descriptions that are related with a verb are solved with a 60% precision.

The failures have been studied and we have found the following reasons:

Failures due to lack of semantic knowledge, since the dictionary of synonyms and relations between names and verbs (which has been made by hand) is quite small (it only has 200 entries). Some cases need more semantic knowledge to be solved. For example, when an antecedent with the same head is chosen but it is not the right solution as it has opposed modifiers, e.g. *the right ear and the left ear*.

The accessibility space used in our algorithm scopes all previous sentences. Therefore, there are a high number of possible antecedents for many definite descriptions. This fact increases the ambiguity and complexity to solve anaphora. Other algorithms such as those developed by Kameyama in [6] only use the ten previous sentences.

[3] It should be remarked that these results cannot be directly compared since we have worked with different corpora and languages.

## 5. Conclusion

An algorithm to classify definite descriptions into anaphoric or non-anaphoric and to solve identity co-references produced by definite descriptions is presented. This algorithm solves co-references produced by definite descriptions with the same head as their antecedent, definite descriptions semantically related with their antecedent and definite descriptions semantically related with the verb (role thematic). It has a precision of 80% on Spanish narrative texts. This algorithm has the following main contributions with reference to other works: it solves definite descriptions that do not have the same head as its antecedent, and do not have a synonym, hyperonym or hyponym relationships, but they have a relationships with the verb of the sentence in which the antecedent appears. Moreover, a classification of definite descriptions based on the different knowledge that is needed to solve the references.

Our future algorithm aims will use other semantic tools such as Spanish WordNet that will provide us with more detailed semantic knowledge. Moreover, in order to solve the problem of the best *accessibility space* (number of sentences in which the solution of an anaphora is scanned), the algorithm will scan a lower number of sentences, and if there is no solution, then a higher number of sentences will be scanned. Then, a study of the best accessibility space will be carried out. Furthermore, this algorithm will be applied on English texts, in order to compare it with other methods.

## 6. Aknowledgements

This paper has been supported by the Comisin Interministerial de Ciencia y Tecnologa, project number TIC97-0671-C02-02.

We are grateful for discussing the central issues contained in this paper to Dr. Manuel Palomar (University of Alicante).

## References

[1] P. Christopherson. *The Articles: A study of their theory and use in English.* E. Munksgaard, Copenhagen, 1939.

[2] H. H. Clark. Bridging. In P. Johnson-Laird and P. Wason, editors, *Thinking: readings in cognitive science*, pages 411–420. Cambridge: CUP, 1977.

| Antecedent type | Total | Failures | Successly solved | Precision |
|---|---|---|---|---|
| Same head | 105 | 15 | 90 | 85.7% |
| Semantic relation | 20 | 5 | 15 | 75% |
| Related with a verb | 5 | 2 | 3 | 60% |
| Total | 135 | 22 | 108 | 80% |

**Table 3. Results obtained applying the algorithm.**


[3] A. Ferrández. *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos*. PhD thesis, Universidad de Alicante, November 1998.

[4] J. A. Hawkins. *Definiteness and indefiniteness*. Humanities Press, Atlantic Highlands, NJ, 1978.

[5] J. R. Hobbs, D. E. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. *In E. Roche and Y. Schabes, eds. Finite State Devices for natural Language Processing. MIT Press, Cambridge, Massachussetts*, 1996.

[6] M. Kameyama. Recognizing Referential Links: An Information Extraction Perspective. In R. In Mitkov and e. B. Boguraev, editors, *Proceedings of ACL / EACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, July 1997.

[7] F. Llopis, R. Muñoz, A. Suárez, and A. Montoyo. EXIT: Propuesta de un sistema de extracción de información de textos notariales. *Revista Nováatica*, 133:26–30, mayo-junio 1998.

[8] P. Martínez-Barco, J. Peral, A. Ferrández, L. Moreno, and M. Palomar. Analizador Parcial SUPP. In *Proceedings of VI biennial Iberoamerican Conference on Artificial Intelligence, IBERAMIA'98*, pages 329–341, 1998.

[9] R. Muñoz, A. Ferrández, and P. Martínez-Barco. Método para la resolución de correferencias de sintagmas nominales definidos incluyendo alias y acrónimos en el sistema de extracción de información EXIT. *Procesamiento del Lenguaje Natural*, 25:143–149, september 1999.

[10] M. Palomar, A. Ferrández, L. Moreno, M. Saiz-Noeda, R. Muñoz, P. Martínez-Barco, J. Peral, and B. Navarro. A Robust Partial Parsing Strategy based on the Slot Unification Grammars. In *Proceeding of 6e Confrence annuelle sur le Traitement Automatique des Langues Naturelles. TALN'99*, pages 263–272, July 1999.

[11] M. Poesio and R. Vieira. A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics. MIT Press*, 24:183–216, 1998.

[12] E. Prince. Toward a taxonomy of given-newinformation. *In P.Cole, editor, Radical Pragmatics. Academic Press, New York*, pages 223–256, 1981.

[13] R. Vieira. Co-reference resolution of definite descriptions. In *Proceedings of VI Simposio Internacional de comunicación Social*, pages 497–503, January 1999.

[14] R. Vieira and M. Poesio. *Corpus-based and computational aproach to anaphora*, chapter Processing definite descriptions in corpora. S.Botley and T. McEnery eds. UCL Press, London, 1998.