An Algorithm for Anaphora Resolution in Spanish Texts

Manuel Palomar*
University of Alicante

Antonio Ferrández*
University of Alicante

Lidia Moreno†
Valencia University of Technology

Patricio Martínez-Barco*
University of Alicante

Jesús Peral*
University of Alicante

Maximiliano Saiz-Noeda*
University of Alicante

Rafael Muñoz*
University of Alicante

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

This paper presents an algorithm for identifying noun phrase antecedents of third person personal pronouns, demonstrative pronouns, reflexive pronouns, and omitted pronouns (zero pronouns) in unrestricted Spanish texts. We define a list of constraints and preferences for different types of pronominal expressions, and we document in detail the importance of each kind of knowledge (lexical, morphological, syntactic, and statistical) in anaphora resolution for Spanish. The paper also provides a definition for syntactic conditions on Spanish NP-pronoun noncoreference using partial parsing. The algorithm has been evaluated on a corpus of 1,677 pronouns and achieved a success rate of 76.8%. We have also implemented four competitive algorithms and tested their performance in a blind evaluation on the same test corpus. This new approach could easily be extended to other languages such as English, Portuguese, Italian, or Japanese.

1. Introduction

We present an algorithm for identifying noun phrase antecedents of personal pronouns, demonstrative pronouns, reflexive pronouns, and omitted pronouns (zero pronouns) in Spanish. The algorithm identifies both intrasentential and intersentential antecedents and is applied to the syntactic analysis generated by the **slot unification parser** (SUP) (Ferrández, Palomar, and Moreno 1998b). It also combines different forms of knowledge by distinguishing between constraints and preferences. Whereas constraints are used as combinations of several kinds of knowledge (lexical, morphological, and syntactic), preferences are defined as a combination of heuristic rules extracted from a study of different corpora.

We present the following main contributions in this paper:

- an algorithm for anaphora resolution in Spanish texts that uses different kinds of knowledge

* Department of Software and Computing Systems, Alicante, Spain. E-mail: (Palomar) mpalomar@dlsi.ua.es, (Ferrández) antonio@dlsi.ua.es, (Martínez-Barco) patricio@dlsi.ua.es, (Peral) jperal@dlsi.ua.es, (Saiz-Noeda) max@dlsi.ua.es, (Muñoz) rafael@dlsi.ua.es
† Department of Information Systems and Computation, Valencia, Spain. E-mail: lmoreno@dsic.upv.es

© 2001 Association for Computational Linguistics

- an exhaustive study of the importance of each kind of knowledge in Spanish anaphora resolution
- a proposal concerning syntactic conditions on NP-pronoun noncoreference in Spanish that can be evaluated on a partial parse tree
- a proposal regarding preferences that are appropriate for resolving anaphora in Spanish and that could easily be extended to other languages
- a blind test of the algorithm
- a comparison with other approaches to anaphora resolution that we have applied to Spanish texts using the same blind test

In Section 2, we show the classification scheme we used to identify the different types of anaphora that we would be resolving. In Section 3, we present the algorithm and discuss its main properties. In Section 4, we evaluate the algorithm. In Section 5, we compare our algorithm with several other approaches to anaphora resolution. Finally, we present our conclusions.

2. Our Classification Scheme for Pronominal Expressions in Spanish

In this section, we present our classification scheme for identifying the different types of anaphora that we will be resolving. Personal pronouns (PP_R), demonstrative pronouns (DP_R), reflexive pronouns (RP_R), and omitted pronouns (OP_R) are some of the most frequent types of anaphoric expressions found in Spanish and are the main subject of this study. Personal and demonstrative pronouns are further classified according to whether they appear within a prepositional phrase (PP) or whether they are complement personal pronouns (clitic pronouns[1]). We present examples for each of the four types of common anaphora. Each example is presented in three forms: as a Spanish sentence, as a word-to-word translation into English, and correctly translated into English.[2]

2.1 Clitic Personal Pronouns (CPP_R)

In the case of clitic personal pronouns, *lo, la, le* 'him, her, it' and *los, las, les* 'them', we consider that the third person personal pronoun plays the role of the complement.

(1) Ana abre [la verja]$_i$ y la$_i$ cierra tras de sí.
Ana opens [the gate]$_i$ and it$_i$ closes after herself
'Ana opens the gate and closes it after herself.'

2.2 Personal Pronouns Not Included in a PP (PP_RnotPP)

We include in this class the personal pronouns *él, ella, ello* 'he, she, it' and *ellas, ellos* 'they'.

(2) Andrés$_i$ es mi vecino. Él$_i$ vive en el segundo piso.
Andrés$_i$ is my neighbor He$_i$ lives on the second floor
'Andrés is my neighbor. He lives on the second floor.'

1 According to Mathews (1997), a clitic pronoun is a pronoun that is treated as an independent word in syntax but that forms a phonological unit with the verb that precedes or follows it.
2 Coindexing indicates coreference between anaphor and antecedent.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

2.3 Personal Pronouns Included in a PP (PP_RinPP)

We include in this class the personal pronouns *él, ella, ello* 'him, her, it' and *ellas, ellos* 'them'.

(3) Juan$_i$ debe asistir pero Pedro lo hará por él$_i$.
Juan$_i$ must attend but Pedro it will do for him$_i$
'Juan must attend but Pedro will do it for him.'

2.4 Demonstrative Pronouns Not Included in a PP (DP_RnotPP)

We include in this class the demonstrative pronouns *éste, ésta, esto* 'this'; *éstos, éstas* 'these'; *ése, ésa, aquél, aquélla* 'that'; and *ésos, ésas, aquéllos, aquéllas* 'those'.

(4) El Ferrari$_i$ ganó al Ford. Éste$_i$ es el mejor.
the Ferrari$_i$ beat the Ford This$_i$ is the best
'The Ferrari beat the Ford. This is the best.'

2.5 Demonstrative Pronouns Included in a PP (DP_RinPP)

We include in this class the demonstrative pronouns *éste, ésta, esto* 'this'; *éstos, éstas* 'these'; *ése, ésa, aquél, aquélla* 'that'; and *ésos, ésas, aquéllos, aquéllas* 'those'.

(5) Ana vive con Paco$_i$ y cocina para éste$_i$ diariamente.
Ana lives with Paco$_i$ and cooks for this$_i$ every day
'Ana lives with Paco and cooks for him every day.'

2.6 Reflexive Pronouns (RP_R)

We include in this class the reflexive pronouns *se, sí, sí mismo* 'himself, herself, itself' and *consigo, consigo mismo* 'themselves'.

(6) Ana$_i$ abre la verja y la cierra tras de sí$_i$.
Ana$_i$ opens the gate and it closes after herself$_i$
'Ana opens the gate and closes it after herself.'

2.7 Omitted Pronouns (Zero Pronouns OP_R)

The omitted pronoun is the most frequent type of anaphoric expression in Spanish, as we will show in Section 4.2. Omitted pronouns occur when the pronominal subject is omitted. This kind of pronoun also occurs in other languages, such as Portuguese or Japanese; in these languages, it can also appear in object position, whereas in Spanish or Italian, it can appear only in subject position. In the following example, the omission is represented by the symbol Ø (the symbol does not appear in the correct translation into English).

(7) Ana$_i$ abre la verja y Ø$_i$ la cierra tras de sí.
Ana$_i$ opens the gate and Ø$_i$ it closes after herself
'Ana opens the gate and she closes it after herself.'

3. Anaphora Resolution Algorithm

In the algorithm, all the types of anaphora are identified from left to right as they appear in the sentence. The most important proposals for anaphora resolution—such as those of Baldwin (1997), Lappin and Leass (1994), Hobbs (1978), or Kennedy and Boguraev (1996)—are based on a separation between **constraints** and **preferences**.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

Constraints discard some of the candidates, whereas preferences simply sort the remaining candidates. A constraint defines a property that must be satisfied in order for any candidate to be considered as a possible solution of the anaphor. For example, pronominal anaphors and antecedents must agree in person, gender, and number.[3] Otherwise, the candidate is discarded as a possible solution. A preference is a characteristic that is not always satisfied by the solution of an anaphor. The application of preferences usually involves the use of heuristic rules in order to obtain a ranked list of candidates.

Each type of anaphora has its own set of constraints and preferences, although they all follow the same general algorithm: constraints are applied first, followed by preferences.

Based on the preceding description, our algorithm contains the following main components:

- identification of the type of pronoun
- constraints
  - morphological agreement (person, gender, and number)
  - syntactic conditions on NP-pronoun noncoreference
- preferences

In order to apply this algorithm to unrestricted texts, it has been necessary to use partial parsing. In our partial-parsing scheme, as presented in Ferrández, Palomar, and Moreno (1999), we only parse coordinated NPs and PPs, verbal chunks, pronouns, and what we have called free conjunctions (i.e., conjunctions that do not join coordinated NPs or PPs). Words that do not appear within these constituents are simply ignored. The NP constituents include coordinated adjectives, relative clauses, coordinated PPs, and appositives as modifiers.

With this partial-parsing scheme, we divide a sentence into clauses by parsing first the free conjunction and then the verbs, as in the following example:

(8)	Pedro	compró	un	regalo	y	se	lo	dio	a	Ana.
	Pedro	bought	a	gift	and	her	it	gave	to	Ana
	'Pedro bought a gift and gave it to Ana.'									

In this example, we have parsed the following constituents: *np(Pedro)*, *v(compró)*, *np(un regalo)*, *freeconj(y)*, *pron(se)*, *pron(lo)*, *v(dio)*, *pp(a Ana)*. We are able to divide this sentence into two clauses because it contains the free conjunction *y* 'and' and the two verbs *compró* 'bought' and *dio* 'gave'.

3.1 Identification of the Kind of Pronoun

The algorithm uses partial-parse trees to automatically identify omitted pronouns by employing the following steps:

- The sentence is divided into clauses (by parsing the free conjunction followed by the verbs).

3 In our implementation, this morphological information is extracted from the part-of-speech tagger.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

- An NP or pronoun is sought for each clause by analyzing the clause constituents on the left-hand side of the verb, unless the verb is imperative or impersonal. The chosen NP or pronoun must agree in person and number with the clausal verb. (In evaluating this algorithm, Ferrández and Peral [2000] achieved a success rate of 88% for detecting omitted pronouns.)

The remaining pronouns are identified based on part-of-speech (POS) tagger outputs.

3.2 Morphological Agreement

Person, gender, and number agreement are checked in order to discard potential antecedents. For example, in the sentence

(9) Juan$_j$ vio a Rosa$_i$. Ella$_i$ estaba muy feliz.
Juan$_j$ saw to Rosa$_i$ She$_i$ was very happy
'Juan saw Rosa. She was very happy.'

there are two possible antecedents for *ella* 'she', whose slot structures[4] are

```
np (conc (sing, masc), X, Juan)
np (conc (sing, fem), Y, Rosa)
```

whereas the slot structure of the pronoun is

```
pron (conc (sing, fem), Z, ella).
```

In order to decide between the two antecedents, the unification of both slot structures (pronoun and candidate) is carried out by the slot unification parser (Ferrández, Palomar, and Moreno 1999). In this example, the candidate *Juan* is rejected by this morphological agreement constraint.

3.3 Syntactic Conditions on NP-Pronoun Noncoreference

These conditions are based on c-command and minimal-governing-category constraints as formulated by Reinhart (1983) and on the noncoreference conditions of Lappin and Leass (1994). They are of great importance in any anaphora resolution system that does not use semantic information, as is the case with our proposal. In such systems, recency is important in selecting the antecedent of an anaphor. That is to say, the closest NP to the anaphor has a better chance of being selected as the solution. One problem, however, is that such constraints are formulated using full parsing, whereas if we want to work with unrestricted texts we should be using partial parsing, as previously defined.

We have therefore proposed a set of noncoreference conditions for Spanish, using partial parsing, although they could easily be extended to other languages such as English. In our system, the following types of pronouns are noncoreferential with a noun phrase (NP) under the conditions noted (noncoindexing indicates that a candidate is rejected by these conditions).

4 The term *slot structure* is defined in Ferrández, Palomar, and Moreno (1998b). The slot structure stores morphological and syntactic information related to the different constituents of a sentence.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

1. Reflexive pronouns are noncoreferential when:

   (a) the NP is included in another constituent (e.g., the NP is included in a PP)

   (10) Ante Luis$_j$ se$_i$ frotó con la toalla.
   in front of Luis$_j$ himself$_i$ rubbed with the towel
   'He rubbed himself with the towel in front of Luis.'

   In this sentence, we would have obtained the following sequence of constituents after our partial-parsing scheme: *pp(prep(ante), np(Luis)), pron(se), v(frotó), pp(prep(con), np(la toalla))*. Following the above-stated condition, the NP *Luis* cannot corefer with the reflexive pronoun *se* since *Luis* is included in a PP (*ante Luis*).

   (b) the NP is in a different clause or sentence

   (11) Ana$_j$ trajo un cuchillo y Eva$_i$ se$_i$ cortó.
   Ana$_j$ brought a knife and Eva$_i$ herself$_i$ cut
   'Ana brought a knife and Eva cut herself.'

   (c) the NP appears after the verb and there is another NP in the same clause before the verb

   (12) Juan$_i$ se$_i$ cortó con el cuchillo$_j$.
   Juan$_i$ himself$_i$ cut with the knife$_j$
   'Juan cut himself with the knife.'

   Under these conditions, coreference is allowed between the NP and the reflexive pronoun, since both are in the same clause. For example:

   (13) Juan$_i$ quería verlo por sí mismo$_i$.
   Juan$_i$ wanted see it for himself$_i$
   'Juan wanted to see it for himself.'

   In this example, *Juan* and the reflexive pronoun *sí mismo* 'himself' corefer since *Juan* is in the same clause as the anaphor, it is not included in another constituent, and it appears before the verb.

2. Clitic pronouns are noncoreferential when:

   (a) the NP is included in a PP (except those headed by the preposition *a* 'to')

   (14) Con Juan$_i$ lo$_j$ compré.
   with Juan$_i$ it$_j$ bought
   'I bought it with Juan.'

   (b) the NP is located more than three constituents before the clitic pronoun in the same clause

   (15) En casa$_i$ [el martillo]$_j$ no se lo$_j$ di.
   at home$_i$ [the hammer]$_j$ not him it$_j$ gave
   'I didn't give him the hammer at home.'

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

In this example, the direct object *el martillo* 'the hammer' has been moved from its common position after the verb, and it is necessary to fill the resulting gap with the pronoun *lo* 'it' even though it does not appear in the English translation. This phenomenon[5] can be considered an exception to the c-command constraints as formulated by Reinhart when applied to Spanish clitic pronouns.

Moreover, if the last two conditions are not fulfilled by the NP and the verb is in the first or second person, then this NP will necessarily be the solution of the pronoun:

(16) [El bolígrafo]$_i$ lo$_i$ comprarás en esa tienda.
[The pen]$_i$ it$_i$ will buy in that shop
'You will buy the pen in that shop.'

3. Personal and demonstrative (nonclitic) pronouns are noncoreferential when the NP is in the same clause as the anaphor, and:

(a) the pronoun comes before the verb (in full parsing, this would mean that it is the subject of its clause)

(17) Ante Luis$_i$ él$_j$ saludó a Pedro$_k$.
in front of Luis$_i$ he$_j$ greeted to Pedro$_k$
'He greeted Pedro in front of Luis.'

(b) the pronoun comes after the verb (in full parsing, this would mean that it is the object of the verb) and the NP is not included in another NP

(18) [El padre de Juan$_j$]$_i$ le venció a él$_j$.
[Juan$_j$'s father]$_i$ him beat to him$_j$
'Juan's father beat him.'

In this example, the pronoun *él* 'him' cannot corefer with the NP *el padre de Juan* 'Juan's father', but it can corefer with *Juan* since it is a modifier of the NP *el padre de Juan*.

It should be mentioned that the clitic pronoun *le* is another form of the pronoun *él* 'him'. This is a typical phenomenon in Spanish, where clitic pronouns occupy the object position. Sometimes both the clitic pronoun and the object appear in the same clause, as occurs in the previous example and in the following one:

(19) A Pedro$_i$ yo le$_i$ vi ayer.
to Pedro$_j$ I him$_i$ saw yesterday
'I saw Pedro yesterday.'

This example also illustrates the previously mentioned exception of c-command constraints for Spanish clitic pronouns. In this case, the direct object *a Pedro* 'to Pedro' has been moved before the verb, and the clitic pronoun *le* 'him' has been added. It should also be remarked that, as noted earlier, the clitic pronoun does not appear in the English translation.

5 Mathews (1997) calls this phenomenon "clitic doubling" and defines it as the use of a clitic pronoun with the same referent and in the same syntactic function as another element in the same clause.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

(c) the pronoun is included in a PP that is not included in another constituent and the NP is not included in another constituent (NP or PP)

(20) [El padre de Luis$_j$]$_i$ juega con él$_j$.
[Luis$_j$'s father]$_i$ plays with him$_j$
'Luis's father plays with him.'

In this example, the pronoun *él* 'him' is included in a PP (which is not included in another constituent) and the NP *el padre de Luis* is not included in another NP or PP. Therefore, the NP cannot corefer with the pronoun. However, the NP *Luis* can corefer because it is included in the NP *el padre de Luis*.

(d) the pronoun is included in an NP, so that the NP in which the pronoun is included cannot corefer with the pronoun

(21) Pedro$_i$ vio [al hermano de él$_i$]$_j$.
Pedro$_i$ saw [the brother of him$_i$]$_j$
'Pedro saw his brother.'

(e) the pronoun is coordinated with other NPs, so that the other coordinated NPs cannot corefer with the pronoun

(22) Juan$_i$, [el tio de Ana]$_j$, y él$_k$ fueron de pesca.
Juan$_i$, [Ana's uncle]$_j$, and he$_k$ went fishing
'He, Juan, and Ana's uncle went fishing.'

(f) the pronoun is included in a relative clause, and the following condition is met:

i. the NP in which the relative clause is included does not corefer with the pronoun

(23) Pedro$_j$ vio a [un amigo que juega con él$_j$]$_i$.
Pedro$_j$ saw to [a friend that plays with him$_j$]$_i$
'Pedro saw a friend that he plays with.'

ii. the NPs that are included in the relative clause follow the previous conditions

iii. the remaining NPs outside the relative clause could corefer with the pronoun

4. Personal and demonstrative (nonclitic) pronouns are noncoreferential when the NP is not in the same clause as the pronoun. (In this case, the NP can corefer with the pronoun, except when this NP also appears in the same sentence and clause as the pronoun, in which case it will have been discarded by the previous noncoreference conditions.)

(24) Ana$_j$ y Eva$_i$ son amigas. Eva$_i$ le$_j$ ayuda mucho.
Ana$_j$ and Eva$_i$ are friends Eva$_i$ her$_j$ helps a lot
'Ana and Eva are friends. Eva helps her a lot.'

It is important to note that the above-mentioned conditions refer to those coordinated NPs and PPs that have been partially parsed. Moreover, as previously mentioned, NPs can include relative clauses, appositives, coordinated PPs, and adjectives.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

We should also remark that we consider a constituent *A* to be included in a constituent *B* if *A* modifies the head of *B*. Let us consider the following NP:

(25) [el hombre que ama a [una mujer que le_i ama$]_j]_i$
[the man who loves to [a woman who him_i loves$]_j]_i$
'the man who loves a woman who loves him.'

We consider that the pronoun *le* 'him' is included in the relative clause that modifies the NP *una mujer que le ama* 'a woman who loves him', which then cannot corefer with it due to noncoreference condition 3(f)i. Under condition 3(f)iii, however, the pronoun *le* 'him' could corefer with the entire NP *el hombre que ama a una mujer que le ama* 'the man who loves a woman who loves him'.

Another example might be the following:

(26) Eva_i tiene [un tío que le_i toma el pelo$]_j$.
Eva_i has [an uncle that her_i teases$]_j$
'Eva has an uncle who teases her.'

In this example, the pronoun is included within the relative clause that modifies *un tío* 'an uncle', and therefore cannot corefer with it. But, following condition 3(f)iii, it can corefer with *Eva*.

3.4 Preferences

To obtain the different sets of preferences, we utilized the training corpus to identify the importance of each kind of knowledge that is used by humans when tracking down the NP antecedent of a pronoun. Our results are shown in Table 1. For our analysis, the antecedents for each pronoun in the text were identified, along with their configurational characteristics with reference to the pronoun. Thus, the table shows how often each configurational characteristic is valid for the solution of a particular pronoun. For example, the solution of a reflexive pronoun is a proper noun 53% of the time. The total number of pronoun occurrences in the study was 575. Thus, we were able to define the different patterns of Spanish pronoun resolution and apply them in order to obtain the evaluation results that are presented in this paper. The order of importance was determined by first sorting the preferences according to the percentage of each configurational characteristic; that is, preferences with higher percentages were applied before those with lower percentages. After several experiments on the training corpus, an optimal order—the one that produced the best performance—was obtained. Since in this evaluation phase we processed texts from different genres and by different authors, we can state that the final set of preferences obtained and their order of application can be used with confidence on any Spanish text.

Based on the results presented in Table 1, we have extracted a set of preferences for each type of anaphora (listed below). We have distinguished between those pronouns that are included within PPs and those that are not. That is because when a pronoun is included in a PP, the preposition of this PP sets a preference.

Preferences of omitted pronouns (OP_R):

1. NPs that are not of time, direction, quantity, or abstract type; that is to say, inanimate candidates are rejected (e.g., *half past ten*, *Market Street*, *three pounds*, or *a thing*)
2. NPs in the same sentence as the omitted pronoun

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

Table 1
Percentage validity of types of pronouns for different configuration characteristics of the training corpus (n = 575).

	CPP_R	RP_R	OP_R	PP_RinPP	DP_RinPP	PP_RnotPP	DP_RnotPP
Intrasentential	66	97	57	70	100	60	75
Intersentential	34	3	43	30	0	40	25
NPSentAnt[a]	9	3	4	16	50	9	38
AntPPin[b]	7	9	14	27	50	20	25
AntProper[c]	57	53	63	35	0	43	0
AntIndef[d]	13	0	7	0	0	6	13
AntRepeat[e]	72	66	79	65	50	71	50
AntWithVerb[f]	14	94	20	24	0	26	25
EqualPP[g]	100	100	100	78	100	97	100
EqualPosVerb[h]	79	84	89	46	0	86	38
BeforeVerb[i]	83	91	89	65	50	86	13
NoTime[j]	100	100	100	100	100	100	100
NoQuantity[k]	100	100	100	100	100	97	100
NoDirection[l]	100	100	100	97	100	100	100
NoAbstract[m]	100	100	100	100	100	100	100
NoCompany[n]	100	100	100	100	100	100	100

a If the NP is included in another NP
b If the NP is included in a PP with the preposition *en* 'in'
c If the NP is a proper noun
d If the NP is an indefinite NP
e If the NP has been repeated more than once in the text
f If the NP has appeared with the verb of the anaphor more than once in the text
g If the NP has appeared in a PP more than once in the text
h If the NP occupies the same position with reference to the verb as the anaphor (before or after)
i If the NP appears before its verb
j If the NP is not a time-type
k If the NP is not a quantity-type
l If the NP is not a direction-type
m If the NP is not an abstract-type
n If the NP is not a company-type

3. NPs that are in the same sentence as the anaphor and are also the solution for another omitted pronoun
4. NPs that are in the previous sentence
5. NPs that are not included in another NP (e.g., when they appear inside a relative clause or appositive)
6. NPs that are not included in a PP or are included in a PP when its preposition is *a* 'to' or *de* 'of'
7. NPs that appear before the verb
8. NPs that have been repeated more than once in the text

Preferences of clitic personal pronouns (CPP_R):

1. NPs that are not of time, direction, quantity, or abstract type
2. NPs that are in the same sentence as the anaphor

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

3. NPs that are in the previous sentence
4. NPs that are not included in another NP (e.g., when they appear inside a clause or appositive)
5. NPs that are not included in a PP or are included in a PP when its preposition is *a* 'to' or *de* 'of'
6. NPs that have appeared with the verb of the anaphor more than once

Preferences of personal and demonstrative pronouns that are included in a PP (PP_RinPP and DP_RinPP):

1. NPs that are not of time, direction, quantity, or abstract type; moreover, in the case of personal pronouns, the NP cannot be a company type
2. NPs that are in the same sentence as the anaphor
3. NPs that are in the previous sentence
4. NPs that are not included in another NP (e.g., when they appear inside a relative clause or appositive)
5. NPs that have been repeated more than once in the text
6. NPs that are included in a PP
7. NPs that occupy the same position (before or after) with respect to the verb as the anaphor

Preferences of personal and demonstrative pronouns that are not included in a PP and of reflexive pronouns (PP_RnotPP, DP_RnotPP, and RP_R):

1. NPs that are not of time, direction, quantity, or abstract type; moreover, in the case of personal pronouns, the NP cannot be a company type
2. NPs that are in the same sentence as the anaphor
3. NPs that are in the previous sentence
4. NPs that are not included in another NP (e.g., when they appear inside a relative clause or appositive)
5. NPs that are not included in a PP or that are included in a PP when its preposition is *a* 'to' or *de* 'of'
6. For the case of personal pronouns (PP_RnotPP), NPs that are not included in a PP with the preposition *en* 'in'
7. NPs that appear before their verbs (i.e., the verb of the sentence in which the NP appears)

3.5 Resolution Procedure

The resolution procedure consists of the following steps:

1. Identify the type of anaphora: pronominal (PP_RinPP or PP_RnotPP), demonstrative (DP_RinPP or DP_RnotPP), reflexive (RP_R), or omitted (OP_R).

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

2. Identify the NP candidate antecedents of a pronoun in order to create a list L. The list created will depend on the type of anaphor and the anaphoric accessibility space (empirically obtained from a deep study of the training corpus) and will be developed according to the following criteria:
   - For pronominal anaphora, demonstrative anaphora, and omitted pronouns, NP candidates will appear in the same sentence as the anaphor and in the four previous sentences.
   - For reflexive anaphora, NP candidates will appear in the same sentence as the anaphor.
3. Apply constraints to L to obtain L_1:
   (a) morphological agreement
   (b) syntactic conditions on NP-pronoun noncoreference
4. If the number of elements of $L_1 = 1$, then the solution is that element.
5. If the number of elements of $L_1 = 0$, then the solution is an exophor.
6. If the number of elements of $L_1 > 1$, then apply preferences to L_1 to obtain L_2. Depending on the type of anaphora, a different set and order of preferences will be applied (see Section 3.4).
7. If the number of elements of $L_2 = 1$, then the solution is that element.
8. If the number of elements of $L_2 > 1$, then apply the following three basic preferences in the order shown until only one candidate remains (these three preferences are common to all the pronouns):
   - NPs most repeated in the text
   - NPs that have appeared most with the verb of the anaphor
   - the first candidate of the remaining list (the closest one to the anaphor)

   After applying these basic preferences, the antecedent is obtained.

4. Empirical Evaluation

4.1 Description of Corpora

We have tested the algorithm on both technical manuals and literary texts. In the first instance, we used a portion of the Spanish edition of the Blue Book corpus.[6] This corpus contains the handbook of the International Telecommunications Union CCITT, published in English, French, and Spanish; it is one of the most important collections of telecommunications texts available and contains 5,000,000 words automatically tagged by the Xerox tagger. In the second instance, the algorithm was tested on Lexesp, a corpus[7] that contains Spanish literary texts from different genres and by different

6 CRATER (Proyecto CRATER 1994–1995) Corpus Resources and Terminology Extraction Project. Project supported by the European Community Commission (DG-XIII). Computational Linguistics Laboratory, Faculty of Philosophy and Fine Arts, Autonomous University of Madrid, Spain.

7 The Lexesp corpus belongs to the project of the same name carried out by the Psychology Department of the University of Oviedo and developed by the Computational Linguistics Group of the University of Barcelona, with the collaboration of the Language Processing Group of the Catalonia University of Technology, Spain.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

Table 2
Pronoun occurrences in two types of texts.

	Total	BB Corpus	Lexesp Corpus
Number of pronoun occurrences in the training corpus	575	123	452
Number of pronoun occurrences in the test corpus	1,677	375	1,302

authors. These texts were mainly obtained from newspapers and were automatically tagged by a different tagger than the one used to tag the Blue Book. The portion of the Lexesp corpus that we processed contained various stories, related by a narrator, and written by different authors. As was the case for the Blue Book corpus, this corpus also contained 5,000,000 words. Since we worked on texts from different genres and by different authors, the applicability of our proposal to other kinds of texts is assured.

We selected a subset of the Blue Book corpus and another subset of the Lexesp corpus, and both were annotated with respect to coreference. One portion of the coreferentially tagged corpus (training corpus) was used for improving the rules for anaphora resolution (constraints and preferences), and another portion was reserved for test data (Table 2).

The annotation phase was accomplished in the following manner: (1) two annotators were selected, (2) an agreement was reached between the annotators with regard to the annotation scheme, (3) each annotator annotated the corpus, and, finally, (4) a reliability test (Carletta et al. 1997) was done on the annotation in order to guarantee the results. The reliability test used the *kappa* statistic that measures agreement between the annotations of two annotators in making judgments about categories. In this way, the annotation is considered a classification task consisting of defining an adequate solution among the candidate list. According to Vieira (1998), the classification task when tagging anaphora resolution can be reduced to a decision about whether each candidate is the solution or not. Thus, two different categories are considered for each anaphor: one for the correct antecedent and another for nonantecedents. Our experimentation showed one correct antecedent among an average of 14.5 possible candidates per anaphor after applying constraints. For computing the *kappa* statistic (k), see Siegel and Castellan (1988).

According to Carletta et al. (1997), a k measurement such as $0.68 < k < 0.8$ allows us to draw encouraging conclusions, and a measurement $k > 0.8$ means there is total reliability between the results of the two annotators. In our tests, we obtained a *kappa* measurement of $k = 0.81$. We therefore consider the annotation obtained for the evaluation to be totally reliable.

4.2 Experimental Work

We conducted a blind test over the entire test corpus of unrestricted Spanish texts by applying the algorithm to the partial syntactic structure generated by the slot unification parser.

Over these corpora, our algorithm attained a success rate for anaphora resolution of 76.8%. We define "success rate" as the number of pronouns successfully resolved, divided by the total number of resolved pronouns. The total number of resolved pronouns was 1,677, including personal, demonstrative, reflexive, and omitted pronouns.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

Table 3
Results of blind test.

	CPP_R	RP_R	OP_R	PP_RinPP	DP_RinPP	PP_RnotPP	DP_RnotPP	Total
Num. of pronoun occurrences	228	80	1,099	107	20	94	49	1,677
Num. of cases correctly resolved	162	74	868	70	17	64	34	1,289
Success rate	71.0%	92.5%	78.9%	65.4%	85.0%	68.0%	69.3%	76.8%

All of them were in the third person, with a noun phrase that appeared before the anaphor as their antecedent. Our algorithm's "recall percentage," defined as the number of pronouns correctly resolved, divided by the total number of pronouns in the text, was therefore 76.8%. A breakdown of success rate results for each kind of pronoun is also shown in Table 3. The pronouns were classified so as to provide the option of applying different kinds of knowledge to resolve each category of pronoun.

One of the factors that affected the results was the complexity of the Lexesp corpus, due mainly to its complex narratives. On average, 16 words per sentence and 27 candidates per anaphor were found in this corpus.

In our experiment, a "successful resolution" occurred if the head of the solution offered by our algorithm was the same as that offered by two human experts. We adopted this definition of "success" because it allowed the system to be totally automatic: solutions given by the annotators were stored in a file and were later automatically compared with the solutions given by our system. Since semantic information was not used at all, PP attachments were not always correctly disambiguated. Hence, at times the differences simply corresponded to different subconstituents.

After the evaluation process, we tested the results in order to identify the limitations of the algorithm with respect to the resolution process. We identified the following:

- There were some mistakes in the POS tagging (causing an error rate of around 3%).
- There were some mistakes in the partial parsing with respect to the identification of complex noun phrases (causing an error rate of around 7%) (Palomar et al. 1999).
- Semantic information was not considered (causing an error rate of around 32%). An example of this type of error can be seen in the following text extracted from the Lexesp corpus:

(27) Recuerdo, por ejemplo, **[un pequeño claro en un bosque en medio de las montañas canadienses]**$_i$, con tres lagunas diminutas que, a causa de los sedimentos del agua, tenían distintos y chocantes colores. Esta rareza había hecho **del sitio**$_i$ un espacio sagrado al que peregrinaron los indios durante siglos y seguramente antes los pobladores paleolíticos. Y eso se notaba.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

Canadá es un país muy hermoso, y **aquél**$_i$ no era, ni mucho menos, el lugar más bello: pero guardaba tranquilamente dentro de sí toda su armonía, como los melocotones guardan dentro de sí el duro hueso.
'I remember, for example, **[a small clearing in the woods in the middle of the Canadian mountains]**$_i$, with three tiny lagoons that, due to the water sediments, had different and astonishing colors. This peculiarity had made **the place**$_i$ into a sacred site, to which the Indians made pilgrimages over the centuries, and surely even the Paleolithic Indians before them. And you could feel it.
'**Canada** is a very beautiful country and **that one**$_i$ was by no means the most beautiful place: but it calmly kept within itself all of its harmony, like peaches that keep the hard seeds within.'

In this text, the demonstrative pronoun *aquél* 'that one' corefers with the antecedent *un pequeño claro en un bosque en medio de las montañas canadienses* 'a small clearing in the woods in the middle of the Canadian mountains', which is also linked to the definite noun phrase *el sitio* 'the place'. Our algorithm identified the proper noun *Canadá*, which is in the same sentence, as the anaphor, since the proper noun could only have been discarded by means of semantic information.

As an example of an anaphor that was correctly resolved by the algorithm, we present the following sentence extracted from the Blue Book corpus. In this case, the antecedent *los sistemas de transmisión analógica* 'the systems of analogue transmission' was correctly chosen for the personal pronoun *ellos* 'them': '

(28) En las conexiones largas o de longitud media, es probable que la fuente principal de ruido de circuito estribe en **[los sistemas de transmisión analógica]**$_i$, ya que en **ellos**$_i$ la potencia de ruido suele ser proporcional a la longitud del circuito.
'In long or medium connections, it is probable that the main source of circuit noise comes from **[the systems of analogue transmission]**$_i$, since in **them**$_i$ the noise capacity is usually proportional to the length of the circuit.'

- The remainder of the errors were due to split antecedents (10%), cataphora (2%), exophora (3%), or exceptions in the application of preferences (43%).

5. Comparison with Other Approaches to Anaphora Resolution

5.1 Anaphora Resolution Approaches

Common among all languages is the fact that the anaphora phenomenon requires similar strategies for its resolution (e.g., pronouns or definite descriptions). All languages employ different kinds of knowledge, but their strategies differ only in the manner by which this knowledge is coordinated. For example, in some strategies just one kind of knowledge becomes the main selector for identifying the antecedent, with other kinds of knowledge being used merely to confirm or reject the proposed antecedent. In such cases, the typical kind of knowledge used as the selector is that of discourse structure. Centering theory, as employed by Strube and Hahn (1999) or Okumura and Tamura (1996), uses this type of approach. Other approaches, however, give equal

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

importance to each kind of knowledge and generally distinguish between constraints and preferences (Baldwin 1997; Lappin and Leass 1994; Carbonell and Brown 1988). Whereas constraints tend to be absolute and therefore discard possible antecedents, preferences tend to be relative and require the use of additional criteria (e.g., the use of heuristics that are not always satisfied by all antecedents). Nakaiwa and Shirai (1996) use this sort of resolution model, which involves the use of semantic and pragmatic constraints, such as constraints based on modal expressions, or constraints based on verbal semantic attributes or conjunctions.

Our approach to anaphora resolution belongs in the latter category, since it combines different kinds of knowledge and no knowledge based on discourse structure is included. We choose to ignore discourse structure because obtaining this kind of knowledge requires not only an understanding of semantics but also knowledge about world affairs and the ability to almost perfectly parse any text under discussion (Azzam, Humphreys, and Gaizauskas 1998).

Still other approaches to anaphora resolution are based either on machine learning techniques (Connolly, Burger, and Day 1994; Yamamoto and Sumita 1998; Paul, Yamamato, and Sumita 1999) or on the principles of uncertainty reasoning (Mitkov 1995).

Computational processing of semantic and domain information is relatively expensive when compared with other kinds of knowledge. Consequently, current anaphora resolution methods rely mainly on constraint and preference heuristics, which employ morpho-syntactic information or shallow semantic analysis (see, for example, Mitkov [1998]). Such approaches have performed notably well. Lappin and Leass (1994) describe an algorithm for pronominal anaphora resolution that achieves a high rate of correct analyses (85%). Their approach, however, operates almost exclusively on syntactic information. More recently, Kennedy and Boguraev (1996) proposed an algorithm for anaphora resolution that is actually a modified and extended version of the one developed by Lappin and Leass (1994). It works from the output of a POS tagger and achieves an accuracy rate of 75%.

There are other approaches based on POS tagger outputs as well. For example, Mitkov and Stys (1997) propose a knowledge-poor approach to resolving pronouns in technical manuals in both English and Polish. The knowledge employed in these approaches is limited to a small noun phrase grammar, a list of terms, and a set of antecedent indicators (definiteness, term preference, lexical reiteration, etc.).

Still other approaches are based on statistical information, including the work of Dagan and Itai (1990, 1991) and Ge, Hale, and Charniak (1998), all of whom present a probabilistic model for pronoun resolution.

We have adopted their ideas and adapted their algorithms to partial parsing and to Spanish texts in order to compare our results with their approaches.

With reference to the differences between English and Spanish anaphora resolution, we have made the following observations:

- Syntactic parallelism has played a more important role in English texts than in Spanish texts, since Spanish sentence structure is more flexible than English sentence structure. Spanish is a free-word-order language and has different syntactic conditions, which increases the difficulty of resolving Spanish pronouns (hence, the greater accuracy rate for English texts).
- A greater number of possible antecedents was observed for Spanish pronouns than for English pronouns, due mainly to the greater average

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

length of Spanish sentences (which also makes the resolution of Spanish pronouns more difficult).

- Spanish pronouns usually bear more morphological information. One result is that this constraint tends to discard more candidates in Spanish than in English.

For comparison purposes, we implemented the following approaches on the same Spanish texts that were tested and described in Section 4.1.

5.2 Hobbs's Algorithm

Hobbs's algorithm (Hobbs 1978) is applied to the surface parse trees of sentences in a text. A surface parse tree represents the grammatical structure of a sentence. By reading the leaves of the parse tree from left to right, the original English sentence is formed. The algorithm parses the tree in a predefined order and searches for a noun phrase of the correct gender and number. Hobbs tested his algorithm for the pronouns *he, she, it*, and *they*, using 100 examples taken from three different sources. Although the algorithm is very simple, it was successful 81.8% of the time.

We implemented a version of Hobbs's algorithm for slot unification grammar for Spanish texts. Since full parsing was not done, our specifications for the algorithm were adjusted, as follows:

- NPs were tested from left to right, as they were parsed in the sentence.
- Afterward, the NPs that were included in an NP (breadth-first) were tested.
- This test was interrupted when an NP agreed in gender and number with the anaphor.

The problems we encountered in implementing Hobbs's algorithm are similar to those found in implementing other approaches: the adaptation to partial parsing, and the inherent difficulty of the Spanish language (i.e., its free-word-order characteristics).

The results of our test of this version of Hobbs's algorithm on the test corpus appear in Table 4.

5.3 Approaches Based on Constraints and Proximity Preference

Our approach has also been compared with the typical baseline approach consisting of constraints and proximity preference; that is, the antecedent that appears closest to the anaphor is chosen from among those that satisfy the constraints. For this comparison, the same constraints that were used previously (i.e., morphological agreement and syntactic conditions) were applied here. Then the antecedent at the head of the list of antecedents was proposed as the solution of the anaphor. These results are also listed in Table 4. As can be seen from the table, success rates were lower than those obtained through the joint application of all the preferences.

5.4 Lappin and Leass's Algorithm

An algorithm for identifying the noun phrase antecedents of third person pronouns and lexical anaphors (reflexive and reciprocal) is presented in Lappin and Leass (1994); this algorithm has exhibited a high rate (85%) of correct analyses in English texts. It relies on measures of salience that are derived from syntactic structures and on simple dynamic models of attentional state to select the antecedent noun phrase of a pronoun from a list of candidates.

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

We have implemented a version of Lappin and Leass's algorithm for Spanish texts. The original formulation of the algorithm proposes a syntactic filter on NP-pronoun coreference. This filter consists of six conditions for NP-pronoun noncoreference within any sentence (Lappin and Leass 1994, page 537). In applying this algorithm to Spanish texts, we changed these conditions so as to capture the appropriate context. As mentioned previously, our algorithm does not have access to full syntactic knowledge. Accordingly, we employed partial parsing over the text in our application of Lappin and Leass's algorithm. The salience parameters were weighted (weight appears in parentheses) and applied in the following way:

- **Sentence recency (100)**: Applied when the NP appeared in the same sentence as the anaphor.
- **Subject emphasis (80)**: Applied when the NP was located before the verb of the clause in which it appeared. This heuristic was necessary because of our algorithm's lack of syntactic knowledge. It should be noted, however, that since Spanish is a nearly free-word-order language and the exchange of subject and object positions within Spanish sentences is common, the heuristic is often invalid. For example, the two Spanish sentences *Pedro compró un regalo* 'Pedro bought a present' and *Un regalo compró Pedro* 'A present bought Pedro' are equivalent to one another and to the English sentence *Pedro bought a present*.
- **Existential emphasis (70)**: In this instance, we applied the parameter in the same way as Lappin and Leass, since the entire NP was fully parsed, which allowed us to tell when it was a definite or an indefinite NP.
- **Accusative emphasis (50)**: Applied when the NP appeared after the verb of the clause in which it appeared and the NP did not appear inside another NP or PP. For example, in the sentence *Pedro encontró el libro de Juana* 'Pedro found Juana's book', a value was assigned to *el libro de Juana* 'Juana's book' but not to *Juana*. Once again, it should be noted that this heuristic was necessary because of our algorithm's lack of syntactic knowledge.
- **Indirect object and oblique complement emphasis (40)**: Applied when the NP appeared in a PP with the Spanish preposition *a* 'to', which usually preceded the indirect object of its sentence.
- **Head noun emphasis (80)**: Applied when the NP was not contained in another NP.
- **Nonadverbial emphasis (50)**: Applied when the NP was not contained in an adverbial PP. In this case, its application depended on the kind of preposition in which the NP was included.
- **Parallelism reward (35):** Applied when the NP occupied the same position as the anaphor with reference to the verb of the sentence (before or after the verb).

Finally, we followed Lappin and Leass in assigning the additional salience value to NPs in the current sentence and in degrading the salience of NPs in preceding sentences.

Our results exhibited some similarities with Lappin and Leass's experiments. For example, anaphora was strongly preferred over cataphora, and both approaches

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

preferred intrasentential NPs to intersentential ones. These results can be seen in Table 4.

5.5 Centering Approach

The centering model proposed by Grosz, Joshi, and Weinstein (1983, 1995) provides a framework for modeling the local coherence of discourse. The model has two constructs, a list of forward-looking centers and a backward-looking center, that can be assigned to each utterance U_i. The list of forward-looking centers $C_f(U_i)$ ranks discourse entities within the utterance U_i. The backward-looking center $C_b(U_{i+1})$ constitutes the most highly ranked element of $C_f(U_i)$ that is finally realized in the next utterance U_{i+1}. In this way, the ranking imposed over $C_f(U_i)$ must reflect the fact that the preferred center $C_p(U_i)$ (i.e., the most highly ranked element of $C_f(U_i)$) is most likely to be $C_b(U_{i+1})$.

The ranking criteria used by Grosz, Joshi, and Weinstein (1995) order items in the C_f list using grammatical roles. Thus, entities with a subject role are preferred to entities with an object role, and objects are preferred to others (adjuncts, etc.).

Grosz, Joshi, and Weinstein (1995) state that if any element of $C_f(U_i)$ is realized by a pronoun in U_{i+1}, then $C_b(U_{i+1})$ must also be realized by a pronoun.

Brennan, Friedman, and Pollard (1987) applied the centering model to pronoun resolution. They based their algorithm on the fact that centering transition relations will hold across adjacent utterances.

Moreover, one crucial point in centering is the ranking of the forward-looking centers. Grosz, Joshi, and Weinstein (1995) state that C_f may be ordered using different factors, but they only use information about grammatical roles. However, both Strube (1998) and Strube and Hahn (1999) point out that it is difficult to define grammatical roles in free-word-order languages like German or Spanish. For languages like these, they propose other ranking criteria dependent upon the information status of discourse entities. They claim that information about familiarity is crucial for the ranking of discourse entities, at least in free-word-order languages.

According to Strube's ranking criteria, two different sets of expressions, *hearer-old* discourse entities (OLD) and *hearer-new* discourse entities (NEW), can be distinguished. OLD discourse entities consist of *evoked* entities—coreferring resolved expressions (pronominal and nominal anaphora, previously mentioned proper names, relative pronouns, appositives)—and *unused* entities (proper names and titles). The remaining entities are assigned to the NEW set. The basic ranking criteria for pronominal anaphora resolution prefer OLD entities over NEW entities.[8]

Strube (1998) thus proposes the following adaptation to the centering model:

- The C_f list is replaced by the list of salient discourse entities (S-list) containing discourse entities that are realized in the current and previous utterance.
- The elements of the S-list are ranked according to the basic ranking criteria and position information:

  If $x \in$ OLD and $y \in$ NEW, then x precedes y.
  If $x, y \in$ OLD or $x, y \in$ NEW,

8 To resolve functional anaphora, a third set, MED, which includes inferable information, must be added between the OLD and the NEW sets. However, this set is not needed to resolve pronominal anaphora (Strube and Hahn 1999).

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

Table 4
Comparative results of blind test.

	Total	CPP_R	RP_R	OP_R	PP_RinPP	DP_RinPP	PP_RnotPP	DP_RnotPP
Num. of pronoun occurrences	1,677	228	80	1,099	107	20	94	49
Hobbs's algorithm	62.7%	61%	85%	62%	62%	50%	66%	52%
Lappin & Leass's algorithm	67.4%	66%	86%	67%	65%	60%	67%	60%
Proximity	52.9%	55%	86%	47%	65%	85%	61%	65%
Centering approach	62.6%	60%	85%	62%	61%	60%	62%	58%
Our algorithm	76.8%	71%	92%	79%	65%	85%	68%	69%

then if $utterance(y)$ precedes $utterance(x)$, then x precedes y,
if $utterance(y) = utterance(x)$ and $pos(x) < pos(y)$, then x precedes y.

- Since there is not a clear definition of what an utterance is, the following criteria are assumed: tensed clauses are defined as utterances on their own and untensed clauses are processed with the main clause in order to constitute only one utterance.

Incorporating these adaptations, Strube (1998) then proposes the following algorithm:

1. If a referring expression is encountered,
   (a) if it is a pronoun, test the elements of the S-list in order until the test succeeds;
   (b) update the S-list using information about this referring expression.
2. If the analysis of utterance U is finished, remove all discourse entities from the S-list that are not realized in U.

The evaluation of this algorithm was performed in Strube (1998) and obtained a precision of 85.4% for English, improving upon the results of the centering algorithm by Brennan, Friedman, and Pollard (1987), which achieved only 72.9% precision when it was applied to the same corpus.

Consequently, in adapting the centering model to Spanish anaphora resolution, we followed Strube's indications. The success rate of the algorithm was not satisfactory, as can be seen in Table 4.

6. Conclusions

In this paper, we have presented an algorithm for identifying noun phrase antecedents of third person personal pronouns, demonstrative pronouns, reflexive pronouns, and

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

omitted pronouns in Spanish. The algorithm is applied to the syntactic structure generated by the slot unification parser—see Ferrández, Palomar, and Moreno (1998a, 1998b, 1999)—and coordinates different kinds of knowledge (lexical, morphological, and syntactic) by distinguishing between constraints and preferences.

The main contribution of this paper is the introduction of an algorithm for anaphora resolution for Spanish. In our work, we have undertaken an exhaustive study of the importance of each kind of knowledge in anaphora resolution for Spanish. Moreover, we have developed a definition of syntactic conditions of NP-pronoun noncoreference in Spanish with partial parsing. We have also adapted our anaphora resolution algorithm to the problem of partial syntactic knowledge, that is to say, when partial parsing of the text is accomplished.

For unrestricted texts, our approach is somewhat less accurate, since semantic information is not taken into account. For such texts, we are dealing with the output of a POS tagger, which does not provide this sort of knowledge. In order to test our approach with texts of different genres by different authors, we have worked with two different Spanish corpora, literary texts (the Lexesp corpus) and technical texts (the Blue Book), containing a total of 1,677 pronoun occurrences.

The algorithm successfully identified the antecedent of the pronoun for 76.8% of these pronoun occurrences. Other algorithms usually work with different kinds of knowledge, different texts, and different languages. In order to make a more valid comparison of our algorithm with others, we adapted the other algorithms so that they would operate using only partial-parsing knowledge. In this evaluation, our algorithm has always obtained better results.

Moreover, based on the results on our study of the importance of each kind of knowledge, we can emphasize that constraints are very important for resolving anaphora successfully, since they considerably reduce the number of possible candidates.

In future studies, we will attempt to evaluate the importance of semantic information in unrestricted texts for anaphora resolution in Spanish texts (Saiz-Noeda, Suárez, and Peral 1999). This information will be obtained from a lexical tool (e.g., Spanish WordNet), which can be automatically consulted (since the tagger does not provide this information).

Acknowledgments

The authors wish to thank Ferran Pla, Natividad Prieto, and Antonio Molina for contributing their tagger (Pla 2000); and Richard Evans, Mikel Forcada, and Rafael Carrasco for their helpful revisions of the ideas presented in this paper. We are also grateful to several anonymous reviewers of *Computational Linguistics* for helpful comments on earlier drafts of this paper. Our work has been supported by the Spanish government (CICYT) with Grant TIC97-0671-C02-01/02.

References

Azzam, Saliha, Kevin Humphreys, and Robert Gaizauskas. 1998. Evaluating a focus-based approach to anaphora resolution. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 74–78, Montreal (Canada).

Baldwin, Breck. 1997. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL/EACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, Madrid (Spain).

Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL'87)*, pages 155–162, Stanford, CA (USA).

Carbonell, Jaime G. and Ralf D. Brown. 1988. Anaphora resolution: A multi-strategy approach. In *Proceedings of the 12th International Conference on*

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

Computational Linguistics (COLING'88), pages 96–101, Budapest (Hungary).
Carletta, Jean, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–32.
Connolly, Dennis, John D. Burger, and David S. Day. 1994. A machine learning approach to anaphoric reference. In *Proceedings of the International Conference on New Methods in Language Processing (NEMLAP'94)*, pages 255–261, Manchester (UK).
Dagan, Ido and Alon Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, pages 330–332, Helsinki (Finland).
Dagan, Ido and Alon Itai. 1991. A statistical filter for resolving pronoun references. In Yishai A. Feldman and Alfred Bruckstein, editors, *Artificial Intelligence and Computer Vision*. Elsevier Science Publishers B. V. (North-Holland), Amsterdam, pages 125–135.
Ferrández, Antonio, Manuel Palomar, and Lidia Moreno. 1998a. A computational approach to pronominal anaphora, one-anaphora and surface count anaphora. In *Proceedings of the Second Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC'98)*, pages 117–128, Lancaster (UK).
Ferrández, Antonio, Manuel Palomar, and Lidia Moreno. 1998b. Anaphora resolution in unrestricted texts with partial parsing. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 385–391, Montreal (Canada).
Ferrández, Antonio, Manuel Palomar, and Lidia Moreno. 1999. An empirical approach to Spanish anaphora resolution. *Machine Translation*, 14(3/4):191–216.
Ferrández, Antonio and Jesús Peral. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, pages 166–172, Hong Kong (China).
Ge, Niyu, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, Montreal (Canada).
Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL'83)*, pages 44–50, Cambridge, MA (USA).
Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
Hobbs, Jerry R. 1978. Resolving pronoun references. *Lingua*, 44:311–338.
Kennedy, Christopher and Branimir Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113–118, Copenhagen (Denmark).
Lappin, Shalom and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.
Mathews, Peter H. 1997. *The Concise Oxford Dictionary of Linguistics*. Oxford University Press, Oxford (UK).
Mitkov, Ruslan. 1995. An uncertainty reasoning approach to anaphora resolution. In *Proceedings of the Natural Language Pacific Rim Symposium (NLPRS '95)*, pages 149–154, Seoul (Korea).
Mitkov, Ruslan. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 869–875, Montreal (Canada).
Mitkov, Ruslan and Malgorzata Stys. 1997. Robust reference resolution with limited knowledge: High precision genre-specific approach for English and Polish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, pages 74–81, Tzigov Chark (Bulgaria).
Nakaiwa, Hiromi and Satoshi Shirai. 1996. Anaphora resolution of Japanese zero pronouns with deictic reference. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 812–817, Copenhagen (Denmark).
Okumura, Manabu and Kouji Tamura. 1996. Zero pronoun resolution in Japanese discourse based on centering theory. In *Proceedings of the 16th International Conference on Computational Linguistics*

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021

(COLING'96), pages 871–876, Copenhagen (Denmark).

Palomar, Manuel, Antonio Ferrández, Lidia Moreno, Maximiliano Saiz-Noeda, Rafael Muñoz, Patricio Martínez-Barco, Jesús Peral, and Borja Navarro. 1999. A robust partial parsing strategy based on the slot unification grammars. In *Proceedings of the 6th Conference on Natural Language Processing (TALN'99)*, pages 263–272, Corsica (France).

Paul, Michael, Kazuhide Yamamoto, and Eiichiro Sumita. 1999. Corpus-based anaphora resolution towards antecedent preference. In *Proceedings of the ACL Workshop on Coreference and Its Applications*, pages 47–52, College Park, MD (USA).

Pla, Ferran. 2000. *Etiquetado Léxico y Análisis Sintáctico Superficial Basado en Modelos Estadísticos*. Ph.D. thesis, Valencia University of Technology, Valencia (Spain).

Proyecto CRATER. 1994-1995. Corpus Resources And Terminology ExtRaction. MLAP-93/20. http://www.lllf.uam.es/proyectos/crater.html (page visited on 04/17/01).

Reinhart, Tanya. 1983. *Anaphora and Semantic Interpretation*. Croom Helm Linguistics series. Croom Helm Ltd., Beckenham, Kent (UK).

Saiz-Noeda, Maximiliano, Armando Suárez, and Jesús Peral. 1999. Propuesta de incorporación de información semántica desde Wordnet al análisis sintáctico parcial orientado a la resolución de la anáfora. *Procesamiento del Lenguaje Natural*, 25:167–173.

Siegel, Sidney and John N. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, NY (USA), 2nd edition.

Strube, Michael. 1998. Never look back: An alternative to centering. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 1251–1257, Montreal (Canada).

Strube, Michael and Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

Vieira, Renata. 1998. *Processing of Definite Descriptions in Unrestricted Texts*. Ph.D. thesis, University of Edinburgh, Edinburgh (UK).

Yamamoto, Kazuhide and Eiichiro Sumita. 1998. Feasibility study for ellipsis resolution in dialogues by machine-learning technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 385–391, Montreal (Canada).

Downloaded from http://direct.mit.edu/coli/article-pdf/27/4/545/1797660/089120101753342662.pdf by guest on 19 June 2021