Coreference-oriented Interlingual Slot Structure & Machine Translation[1]

Jesús Peral, Manuel Palomar and **Antonio Ferrández**
Research Group on Language Processing and Information Systems.
Department of Software and Computing Systems. University of Alicante
03690 San Vicente del Raspeig. Alicante, Spain,
{jperal, mpalomar, antonio}@dlsi.ua.es

Abstract

One of the main problems of many commercial Machine Translation (MT) and experimental systems is that they do not carry out a correct pronominal anaphora generation. As mentioned in Mitkov (1996), solving the anaphora and extracting the antecedent are key issues in a correct translation.

In this paper, we propose an Interlingual mechanism that we have called *Interlingual Slot Structure* (*ISS*) based on *Slot Structure* (*SS*) presented in Ferrández *et al.* (1997). The *SS* stores the lexical, syntactic, morphologic and semantic information of every constituent of the grammar. The mechanism *ISS* allows us to translate pronouns between different languages. In this paper, we have proposed and evaluated *ISS* for the translation between Spanish and English languages. We have compared pronominal anaphora resolution both in English and Spanish to accomplish a study of the existing discrepancies between two languages.

This mechanism could be added to a MT system such as an additional module to solve anaphora generation problem.

Introduction

According to Mitkov (1996), the establishment of the antecedents of anaphora is of crucial importance for a correct translation. It is essential to solve the anaphoric relation when a language is translated into one that marks the pronoun gender. On the other hand, anaphora resolution is vital when translating discourse rather than isolated sentences since the anaphoric references to preceding discourse entities have to be identified. Unfortunately, the majority of Machine Translation (MT) systems do not deal with anaphora resolution and their successful operation usually does not go beyond the sentence level.

Another important aspect in automatic translation of pronouns, as mentioned in Mitkov (1996), consists on the application of two possible techniques: translation or reconstruction of referential expressions.

In the first technique, source language pronouns are directly translated into target language pronouns without studying their relation with other words in the text.

The second technique considers that the pronouns are not autonomous in their meaning/function but dependent on other units in the text. Then, a more natural way to treat pronouns in MT would be the following: (a) analysis has to determine the reference structure of the source text, i.e. coreference/ cospecification relationships between anaphora and antecedents have to be determined, (b) this is the only information that is conveyed to the target language generator, (c) the target language generator generates the appropriate target language surface expression as a function of the target equivalent of the source antecedent and/or according to the rules of this language. Mitkov *et al.* (1995) adopt a similar approach.

In this work, we present an Interlingual (formal language without ambiguity) mechanism proposal based on the second technique. Basically, a structure that stores the anaphora and its antecedent in the source language is used. From this structure, a similar one in the target language is generated. Using this new structure we will be able to generate the final surface structure of the original sentence.

[1] This paper has been supported by the CICYT number TIC97-0671-C02-02.

In the following section we will describe the general purpose anaphora resolution system. The following section will show the anaphora resolution module, where we will focus on the differences between English and Spanish system and we will report some evaluation results. After that, we will present our Interlingual mechanism based on the English-Spanish discrepancy analysis. Finally, we will discuss the evaluation of some commercial MT systems with their problems in pronouns translation and we will study the solution with our proposal.

1 General purpose anaphora resolution system

The general purpose anaphora resolution system with our Interlingual module is shown in Figure 1. It can be observed that there are two processes in parallel, corresponding to anaphora resolution in English and Spanish. These two processes are independent of each other and they are connected by means of the Interlingual mechanism. The input of each process is a grammar defined by means of the grammatical formalism *SUG* (*Slot Unification Grammar*) Ferrández *et al.* (1997), Ferrández (1998a). A translator which transforms rules SUG into Prolog clauses has been developed. This translator will provide a Prolog program that will parse each sentence. This system can carry out a partial or full parsing of the text with the same parser and grammar. In this paper we will use a partial parsing (*Slot Unification Partial Parser*, *SUPP*).
This partial parser *SUPP* described in Martínez-Barco *et al.* (1998), works on unrestricted corpus that contains the words tagged with their obtained grammatical categories from the output of a "part-of-speech (POS) tagger". In this paper, we have used bilingual corpus (*Blue Book*, English and Spanish) CRATER (1994) for the evaluation of anaphora resolution module.
The output of the parsing module will be what we have called *Slot Structure* (henceforth *SS*) that stores the necessary information for linguistic phenomena resolution. This *SS* will be the input for the following module in which we deal with anaphora resolution as well as other linguistic phenomena (extraposition, ellipsis, ...).
After applying the linguistic phenomena resolution algorithm we obtain a new slot structure (*SS'*) that will store both the anaphora and their antecedents. This new structure in the source language will be the input for the Interlingual mechanism (*Interlingual Slot Structure*, *ISS*), which will obtain the corresponding slot structure in the target language. Using this new structure we will be able to generate the final surface structure of the original sentence.

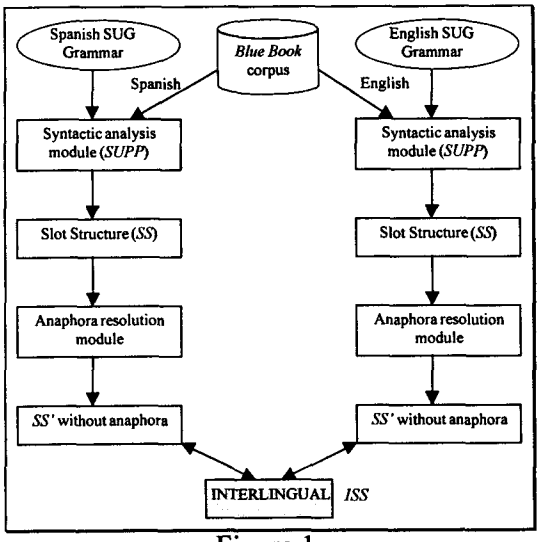

Figure 1.

2 The anaphora resolution module

In this section we will describe the anaphora resolution module of our system. This section consists of two subsections. In the first one we show the algorithm for anaphora resolution. Next, we show the evaluation of the module.

2.1 The algorithm

We are going to describe an algorithm that deals with discourse anaphora in unrestricted texts using partial or full parsing. It is based on the process described in Figure 1. So, this process will be applied after the parsing of a sentence.
This algorithm is shown in Figure 2 and it can deal with pronominal anaphora, surface-count anaphora and one-anaphora as is shown in Ferrández (1998a). This algorithm will use a *Slot Structure* (*SS*) corresponding to the output of the parsing module and a list of antecedents. This list consists of the slot structures of all the previously parsed noun phrases. For each anaphor in this *SS*, several constraints and preferences will be applied. The output of this algorithm consists on a

new SS (*SS'*), where each anaphor has been stored with its correct antecedent.

```
Parse a sentence. We obtain its slot structure (SS1).
For each anaphor in SS1:
  Select the antecedents of the previous X sentences
  depending on the kind of anaphor in L0
  Apply constraints (depending on the kind of anaphor) to L0
  with a result of L1:
  Case of:
    |L1| = 1 Then:
      This one will be the antecedent of the anaphor
    |L1| > 1 Then:
      Apply preferences (depending on the kind of anaphor) to
      L1, with a result of L2:
      The first one of L2 will be the selected antecedent
Update SS1 with each antecedent of each anaphor with a result of SS2.
```

Figure 2.

The detection of the anaphors and possible antecedents is easily carried out by means of the information stored in each *SS*, i.e. its functor and arity. For example, the antecedents have an *SS* with *np* as their functor, whereas the pronouns have *pron*. We have considered the previous two sentences to search for antecedents of a pronoun.

The algorithm will apply a set of constraints (morphosyntactic agreement and c-command constraints) to the list of possible antecedents in order to discount candidates. If there is only one candidate, this one will be the antecedent of the anaphor. Otherwise, if there is still more than one candidate left, a set of preferences (syntactic parallelism, lexical information, reiteration of an antecedent in the text, ...) will be applied. These preferences will sort the list of remaining antecedents, and the first one will be the selected antecedent. These constraints and preferences are described in more detail in Ferrández (1998a), Ferrández *et al.* (1998b).

2.2 Evaluation of the anaphora resolution module

As we reported in Ferrández et. al (1998b), we run our system on part of the Spanish version of *The Blue Book* corpus. We did not use semantic information since the tagger did not provide this information, but in spite of this being lacking we obtained the following figures: it detected 100% of the pronominal anaphors, medium length of sentences with anaphors was 48 words and for pronominal anaphora we obtained 83% accuracy (*pronouns rightly solved divided by the total number of pronouns*). For *The Blue Book* in English, we have obtained the following figures: 79 pronouns (*it*[2]*:41, they:29, themselves:9*) with an accuracy of 87.3% (*it:80,5%, they:93,1%, themselves:100%*); on average, 22 words per sentence.

The reason why some of the references failed is mainly due to the lack of semantic information and due to some weakness of the English grammar that we use. For example, in the sentence (1), our system has not selected the right antecedent (*The French term "communication" and the Spanish term "comunicación"*) due to the symbol " (inverted commas) has been tagged as a new word, and in our grammar we have not foreseen this in a *np*, so the coordination of both *np* have failed.

(1) Note 2 - ***The French term " communication " and the Spanish term " comunicación "*** *have the current meaning given in this definition, but* ***they*** *also acquire a more specific meaning in telecommunication (see 0009 , 0010 and 0011) .*

With reference to the differences between English and Spanish pronoun resolution, we have observed that there is a greater number of possible antecedents for Spanish pronouns (26) than for English (11). This fact could be due to the larger size of Spanish sentences.

Another difference is that constraints (c-command and morphologic agreement) have played a more important role for Spanish texts in the detection of the antecedent: the total number of possible antecedents is reduced from 733 to 222 (a reduction of 70%), whereas for English texts it has only a reduction of 37.7%. This fact is mainly due to the fact that Spanish language has more morphologic information than English.

With regard to the importance of each kind of information for each language, if we apply exactly the same set of preferences in Spanish and English, we obtain a 76% accuracy in English. But we have obtained a better accuracy (87.3%) if we give more importance to syntactic parallelism and less importance to statistical information.

3 Interlingual mechanism focused on MT: Discrepancy analysis

In this section, we will present the Interlingual mechanism *ISS* that takes as input *SS'* (the final

[2] Pleonastic pronouns *it* (i.e. non-anaphoric *it*) have not been included in these results.

slot structure obtained after applying the anaphora resolution module) from the source language and generates the slot structure in the target language. In our proposal, we will study pronominal anaphora generation exclusively. We will divide the section in several subsections that solve the different discrepancies between English and Spanish. In Figure 3 we can see the *ISS*.

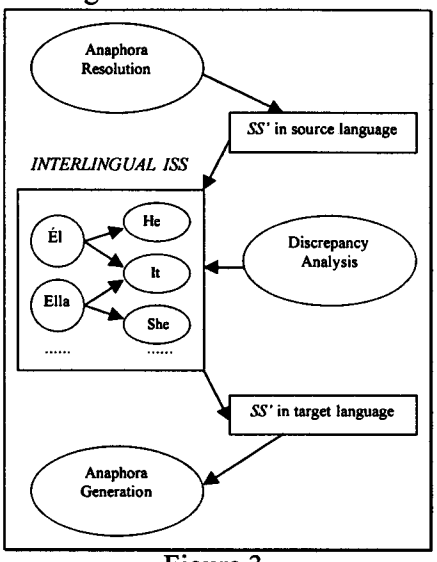

Figure 3.

3.1 Number discrepancy resolution

One problem is generated by the discrepancy between words of different languages that express the same concept. These words can be referred to a singular pronoun in the source language and to a plural pronoun in the target language.

We construct a table with the words that refer to a singular pronoun in the source language and they refer to a plural pronoun in the target language in order to be able to solve these discrepancies correctly. Firstly, we consult this table in the anaphora translation. If the pronoun and its antecedent appear in this figure, we will carry out the indicated transformation.

Anteced	*Span. Anaphor*	*Eng. Anaphor*	*Anteced*
Policía	Ésta	They	Police
Gente	Ésta	They	People
Público	Éste	They	Public
Juventud	Ésta	They	Youth
Ganado	Éste	They	Cattle
Gente	Ésta	They	Folk
...			

Figure 4.

In Figure 4, some examples of these words are shown.

In Figure 5 the English-Spanish translation of a sentence with number discrepancies is described. In this figure, the translation of English *SS*[3] into Spanish *SS* is shown

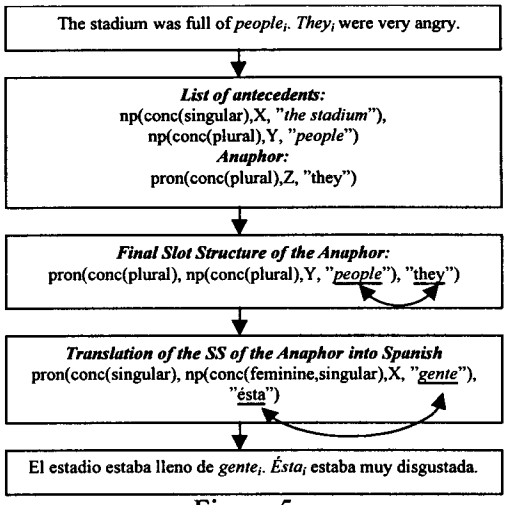

Figure 5.

This *SS* stores for each constituent the following information: constituent name, semantic and morphologic information (structure with functor *conc*), discourse marker (identifier of the entity or discourse object) and the *SS* of its subconstituents. As can be observed in Figure 5 we store in the *SS* of pronouns the information of the right antecedent obtained after applying the anaphora resolution module.

It is necessary to emphasise that after carrying out the translation, the anaphor must agree in number and person with the verb of the sentence where it appears.

3.2 Gender discrepancy resolution

In order to solve personal pronoun gender discrepancies, we construct a table that translates Spanish personal pronouns into the English ones and vice versa.

In the Spanish-English translation we only have problems with the pronoun *it*. The Spanish pronoun él/éste (masculine singular third person) can be translated into *he* or *it*. If the antecedent of the pronoun *él/éste* refers to a person, we will translate it into *he*. If the antecedent of the

[3] Henceforth, we will write the simplified *SS* where it solely appears the relevant information for each example.

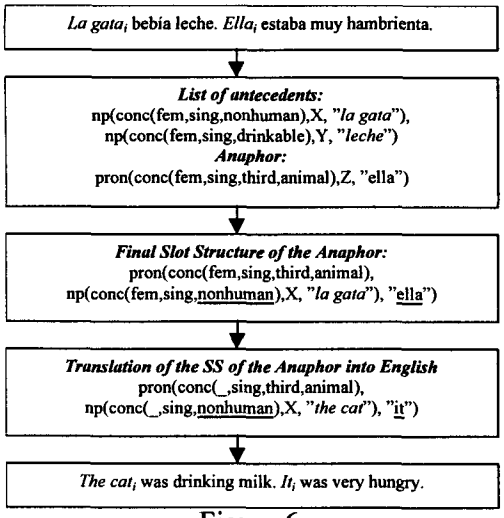

Figure 6.

pronoun is an animal or a thing we will translate it into *it*. These characteristics of the antecedent can be obtained from the semantic information that it is stored in its *SS*. This semantic information can be incorporated to the system using IRSAS method Moreno *et al.* (1992) or another linguistic resource, like WordNet. A similar trouble occurs with the Spanish pronoun *ella/ésta* which is solved in the same way.

In the example of Figure 6 the third argument of the *conc* structures of these *SS* is the semantic type, according to the IRSAS ontology. As it can be observed, the *np "the cat"* has the semantic type nonhuman(animal) and for this reason the pronoun *ella* is translated into the English pronoun *it*.

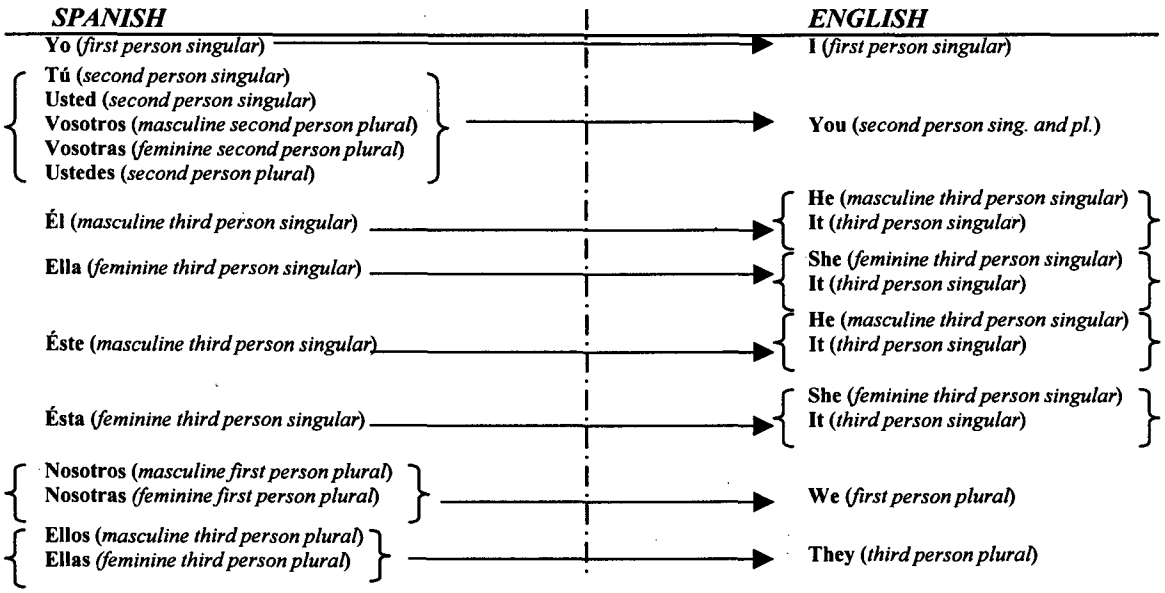

Figure 7.

The table of Figure 7 is used for the remaining pronouns and a direct conversion into English is made.

We have analysed that Spanish has more morphologic information than English, which is extremely relevant in the English-Spanish translation. In order to solve this problem and to choose the right Spanish pronoun we must obtain the gender and number information from the antecedent of the anaphora and carry out the translation. The pronoun *it* involves a series of problems since it can be translated into four different Spanish pronouns (*él, ella, éste, ésta*). These Spanish pronouns refer to both animals and things, but normally *él/ella* refers to animals and *éste/ésta* refers to things. Therefore, in our automatic Interlingual mechanism, when the antecedent of the pronoun is an animal it is translated into *él/ella* and when it is a thing it is translated into *éste/ésta*, since it is the most common use in Spanish.

Finally, an additional difficulty exists in the translation of the pronoun *you*. In Spanish, there are two pronouns for the singular second person (*tú* or *usted*) and three pronouns for the plural second person (*vosotros/vosotras* or *ustedes*). Basically, the difference lies on which the pronouns *tú/vosotros/vosotras* are used in an

informal language (colloquial) whereas *usted/ustedes* are used in a formal one. This implies that to have a specific knowledge of the situation is necessary to be able to choose the right pronoun. Our proposal does not carry out word sense disambiguation and, simply, the colloquial pronouns *tú/vosotros/vosotras* will be chosen in these cases.

3.3 Syntactic discrepancy resolution

This discrepancy is due to the fact that the surface structures of the Spanish sentences are more flexible than the English ones. The constituents of the Spanish sentences can appear in any position of the sentence. In order to carry out a correct translation into English, we must firstly reorganise the Spanish sentence. Nevertheless, in the English-Spanish translation, in general, this reorganisation is not necessary and a direct translation can be carried out.

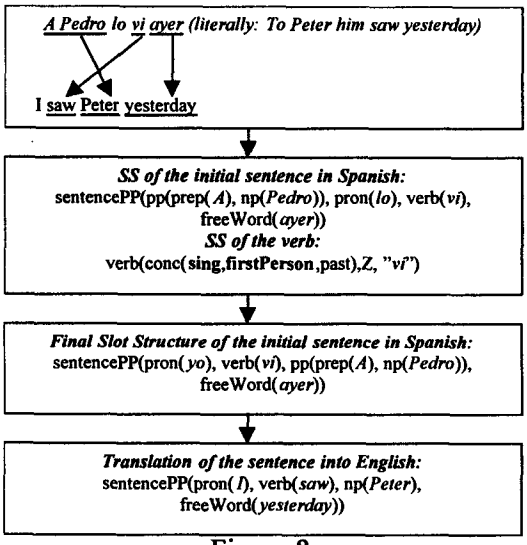

Figure 8.

Let us see an example with the Spanish sentence "*A Pedro lo vi ayer*" (*I saw Peter yesterday*). In this sentence, the object of the verb appears before the verb (in the position of the theoretically subject) and the subject is omitted. Moreover, there is a pronoun, *lo (him)* that functions as complement of the verb *vi (saw)*. This pronoun in Spanish refers to the object of the verb, *Pedro (Peter)*, when it is moved from its theoretical place after the verb (as it occurs in this sentence). In this sentence, the pronominal subject has been omitted. We can find out the subject since the verb is in first person and singular (information stored into its *conc* structure), so the subject would be the pronoun *yo (I)*. Therefore, the solution would be a new *SS* in which the order of the constituents is the usual in English: *subject, verb, complements of the verb*.

In Figure 8, we can see this process graphically. In this sentence, the *pp ("a Pedro")* functions as a indirect object of the verb (because it has the preposition *a (to)*), and the subject of the verb has to be in first person and singular. After reorganising the sentence, we carry out the translation of each constituent. The words that have not been parsed (*freeWord*) are translated into the appropriate words in the target language.

3.4 Elliptical zero-subject construction resolution

Omitting the pronominal subject is usual in Spanish. In these cases, we get the number and person information from the verb to obtain the corresponding English pronoun.

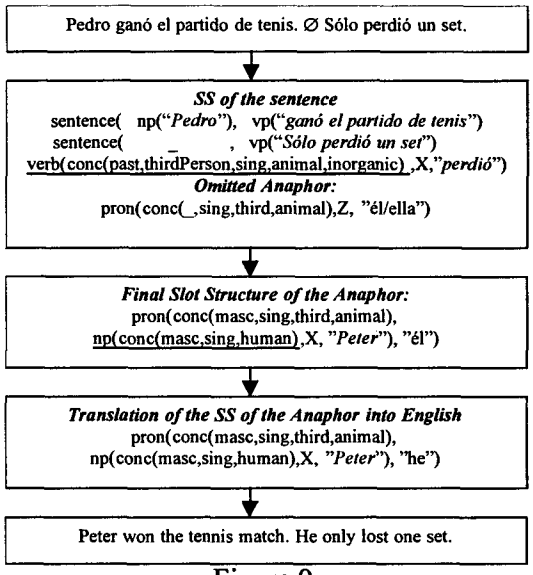

Figure 9.

We can check the omission of the pronominal subject of a sentence by means of the *SS* of the sentence as it is shown in Figure 9. In this figure, we know that the subject of the sentence has been omitted due to the Prolog variable that we find. When it is omitted in the sentence, the *SS* would have a Prolog variable in the slot corresponding to this noun phrase. We can obtain the information corresponding to the subject from the verb of the sentence. In this figure, it would be third person,

singular and masculine or feminine. With these omitted pronominal anaphors, we will apply the preference for the subject of the previous sentence (if it agrees in person and number, and if it is semantically consistent). This information is used to find its antecedent, in this case *Pedro (Peter)* with masculine gender, so the final translation would choose a masculine pronoun (*he*).

Sometimes, we can also obtain the gender information of the pronoun when the verb is copulative. For example, in[4]: *Pedro$_i$ vio a Ana$_j$ en el parque. ∅$_j$ Estaba muy guapa (Peter$_i$ saw Ann$_j$ in the park. **She$_j$** was very beautiful)*. In this example, the verb *estaba (was)* is copulative, so its subject has to agree in gender and number with its object. In this way, we can obtain the gender information from the object, *guapa (beautiful woman)*, that has feminine gender, so the omitted pronoun would be *she* instead of *he*.

4 Commercial MT system evaluation and discussion

In this section, we evaluate different commercial MT systems analysing their deficiencies in translating pronominal anaphora. We study how MT systems deal with the presented discrepancies. In this paper we evaluate 4 systems: (1) Key Translator Pro Version 2.0 (Softkey International), (2) Power Translator Professional (Globalink, Inc.), (3) SYSTRAN Translation Software (http://babelfish.altavista.com/cgi-bin/translate) and (4) DosAmigos version 4.0 (Worldwide Sales Corp.).

In Figure 10, it can be observed the translation of an English-Spanish sentence with gender discrepancies. In (1) and (2) the pronoun *they* is wrongly translated into *ellos* (masculine plural); in (3) and (4) the pronoun is omitted. The pronominal subject can be omitted in Spanish. However, pronominal anaphora is always presented in Spanish in our automatic *ISS* mechanism.

The correct translation of this anaphoric expression in our system is the pronoun *ellas* (feminine plural). The information related to the gender and number must be extracted from the correct antecedent.

Source language: Women were in the duty-free shop. ***They*** were buying gifts for their husbands.	
(1)	Mujeres sido en el exento de derechos de aduana tienda. ***Ellos*** estaban regalos comprantes para sus esposos.
(2)	Las mujeres estaban en la tienda libre de impuestos. ***Ellos*** compraban los regalos para sus esposos.
(3)	Las mujeres estaban en el departamento con franquicia. ∅ Compraban regalos para sus maridos.
(4)	Las mujeres estuvieron en la tienda de libre-de-impuestos. ∅ Estuvieron comprando regalos para sus maridos.
Target language: Las mujeres estaban en la tienda libre de impuestos. ***Ellas*** estaban comprando regalos para sus maridos.	

Figure 10.

In figure 11, an English-Spanish translation with gender discrepancies can be observed. The Spanish pronoun *él* is translated into *he* in (1) (2) (3) and (4) while the right translation is the pronoun *it*. In our proposal, we solve the problem using semantic information of the antecedent. In this case, the antecedent *el mono* (*the monkey*) is an animal, therefore, the pronoun *he* must be translated into *it*.

In figure 12, a number discrepancy can be observed. The word *police* is plural in English, while it is singular in Spanish (*policía*). In (1) (2) (3) and (4) we can observed wrong translations and pronouns that do not agree with the verb. Before the translation, the number discrepancy table is consulted and if the pronoun and its antecedent appear in this table, we will carry out the indicated transformation. After the translation, the anaphor must agree in number and person with the verb of the sentence where it appears.

Source language: El mono se bebió la leche. Después, ***él*** saltó entre los árboles.	
(1)	The monkey was drunk the milk. Afterwards, ***he*** jumped between the trees.
(2)	The monkey was drunk the milk. After, ***he*** jumped between the trees.
(3)	The monkey drank milk. Later, ***he*** jumped between the trees.
(4)	The monkey [bebió] milk her/you/it. [Después], [él] [saltó] ***I/he/she/you*** enter the [árboles].
Target language: The monkey drank milk. Later, ***it*** jumped between the trees.	

Figure 11.

[4] The symbol ∅ in a position of the sentence marks the omitted words in that position.

Source language: The police are coming. ***They*** are just in time.	
(1)	La policía viene. ***Ellos*** son solamente en tiempo.
(2)	Los policías vienen. ***Ellos*** son simplemente en el tiempo.
(3)	El policía está viniendo. ***Él*** es justa en tiempo.
(4)	La policía están viniendo. Ø Justamente son a tiempo.
Target language: La policía está viniendo. ***Ésta*** llegará a tiempo.	

Figure 12.

In Figure 13, an example of Spanish-English syntactic discrepancies can be observed. The systems (1) (2) (3) and (4) fail in the translation. In our mechanism, we reorganise the sentence and then, we accomplish the translation.

Source language: A Pedro lo vi ayer.	
(1)	To I Ask for was seen it yesterday.
(2)	To Pedro I saw it yesterday.
(3)	To Pedro I saw yesterday.
(4)	TO/AT Pedro saw him/you/it yesterday.
Target language: I saw Peter yesterday.	

Figure 13.

Finally, we analyse the Spanish elliptical zero-subject construction. In Figure 14, the systems (1) (2) and (4) fail in the translation. In our proposal, we obtain the information corresponding to the subject from the verb of the sentence. In this example, the pronoun must be first or third person and singular. We extract the gender information from the correct antecedent (feminine) and we determine that the pronoun is *she* (*ella*), feminine third person singular.

Source language: La mujer tenía hambre. Ø Comía el melón.	
(1)	The woman was hungry. Ø Was eatting the melon.
(2)	The woman were hungry. Ø Was eating the melon.
(3)	The woman was hungry. ***She*** ate the melon.
(4)	The woman was being hungry. ***I/he/she/you*** was eating the melon.
Target language: The woman was hungry. ***She*** ate the melon.	

Figure 14.

Conclusion

After the evaluation, we consider that most of the MT systems do not deal with anaphora resolution and their successful operation usually does not go beyond the sentence level. We propose an Interlingual mechanism that relate pronouns in different languages (English-Spanish) with the information stored of the resolution of its antecedent allowing us a correct translation between both languages.

The evaluation of the pronoun translation has been analysed by hand, where we have obtained that if the pronoun resolution is correct, its translation as well. However, we have obtained in pronominal anaphora resolution: 83% and 87.3% accuracy for Spanish and English respectively.

References

CRATER (1994) *Corpus Resources and Terminology Extraction Project.* Proyecto financiado por la Comisión de las Comunidades Europeas (DG-XIII). Investigadores principales Marcos, F. y Sánchez, F. Laboratorio de Lingüística Informática, Facultad de Filosofía y Letras, Univ. Autónoma de Madrid.

Ferrández, A., Palomar, M. and Moreno, L. (1997) Slot Unification Grammar. In *Proceedings of the Joint Conf. on Declarative Programming, APPIA-GULP-PRODE'97* (Grado, Italy, June 1997). pp. 523-532.

Ferrández, A. (1998a) *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos.* Ph.D. Thesis. Dpt. of Lenguajes y Sistemas Informáticos. Univ. of Alicante, Spain, July 1998.

Ferrández A., Palomar M. and Moreno L. (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* (Montreal, Canada, August 1998). pp. 385-391.

Martínez-Barco, P., Peral, J., Ferrández, A., Moreno, L. and Palomar, M. (1998) Analizador Parcial SUPP. In *Proceedings of VI biennial Iberoamerican Conference on Artificial Intelligence, IBERAMIA'98* (Lisbon, Portugal, October 1998). pp. 329-341.

Mitkov R., Choi S.K. and Sharp R. (1995) Anaphora resolution in Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI'95* (Leuven, Belgium, July 1995).

Mitkov, R. (1996) Anaphora and machine translation. Tech. Report. *Machine Translation Review* (1996).

Moreno, L., Andrés, F. and Palomar, M. (1992) Incorporar Restricciones Semánticas en el Análisis Sintáctico: IRSAS. *Procesamiento del Lenguaje Natural, 12* (1992).