An Application of the Interlingua System ISS for Spanish-English Pronominal Anaphora Generation[1]

Jesús Peral 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, antonio}@dlsi.ua.es

Abstract

In this paper, we present the Interlingua system *ISS* to generate the pronominal anaphora into the Spanish and English languages. We also describe the main problems in the pronoun generation into both languages such as zero-subject constructions and number, gender and syntactic differences. Our system improves other proposals presented so far due to the fact that we are able to solve and generate intersentential anaphora, to detect coreference chains and to generate Spanish zero-pronouns into English, issues that are hardly considered by other systems. Finally, we provide outstanding results of our system on unrestricted corpora.

Introduction

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. Solving the anaphora and extracting the antecedent are key issues in a correct generation into the target language. Unfortunately, the majority of MT systems do not deal with anaphora resolution and their successful operation usually does not go beyond the sentence level. In this paper, we present a complete approach that allows pronoun resolution and generation into the target language.

Our approach works on unrestricted texts unlike other systems, like the KANT interlingua system (Leavitt *et al.* (1994)), that are designed for well-defined domains. Although full parsing of these texts could be applied, we have used partial parsing of the texts due to the unavoidable incompleteness of the grammar. This is a main difference with the majority of the interlingua systems such as the DLT system based on a modification of Esperanto (Witkam (1983)), the Rosetta system which is experimenting with Montague semantics as the basis for an interlingua (Appelo and Landsbergen (1986)), the KANT system, etc. as they use full parsing of the text.

After the parsing and solving pronominal anaphora, an interlingua representation of the whole text is obtained. In the interlingua representation no semantic information is used as input, unlike some approaches that have as input semantic information of the constituents (Miyoshi *et al.* (1997), Castellón *et al.* (1998), the DLT system, etc).

From this interlingua representation, the generation of anaphora (including intersentential anaphora), the detection of coreference chains of the whole text and the generation of Spanish zero-pronouns into English have been carried out, issues that are hardly considered by other systems. Furthermore, this approach can be used for other different applications, e.g. Information Retrieval, Summarization, etc.

The paper is organized as follows: In section 1, the complete approach that includes Analysis, Interlingua and Generation modules will be described. These modules will be explained in detail in the next three sections. In section 5, the Generation module has been evaluated in order

[1] This paper has been partly financed by the collaborative research project between Spain and The United Kingdom number HB1998-0068.

to measure the efficiency of our proposal. To do so, two experiments have been accomplished: the generation of Spanish zero-pronouns into English (syntactic generation module) and the generation of English pronouns into Spanish ones (morphological generation module). Finally, the conclusions of this work will be presented.

1 System Architecture

The complete approach that solves and generates the anaphor is based on the scheme of *Figure 1*. Translation is carried out in two stages: from the source language to the Interlingua, and from the Interlingua into the target language. Modules for analysis are independent from modules for generation. In this paper, although we have only studied the Spanish and English languages, our approach is easily extended to other languages, i.e. multilingual system, in the sense that any analysis module can be linked to any generation module.

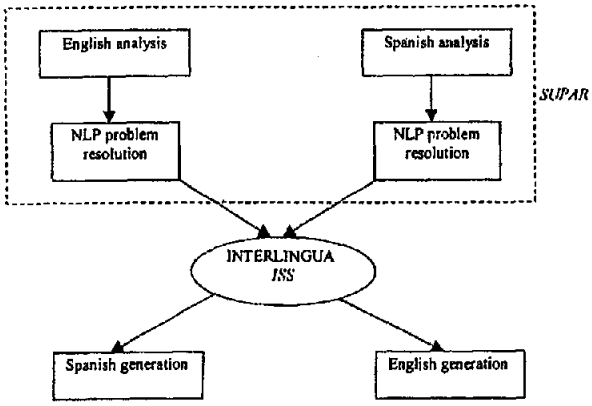

Figure 1. System architecture.

As can be observed in *Figure 1*, there are three independent modules in the process of generation: Analysis, Interlingua and Generation modules.

2 Analysis module

The analysis is carried out by means of *SUPAR (Slot Unification Parser for Anaphora resolution)* system, presented in Ferrández *et al.* (2000). *SUPAR* is a computational system focused on anaphora resolution. It can deal with several kinds of anaphora, such as pronominal anaphora, one-anaphora, surface-count anaphora and definite descriptions. In this paper, we focus on pronominal anaphora resolution and generation into the target language. In pronominal anaphora resolution in both the Spanish and English languages, the system has achieved an accuracy of 84% and 87% respectively.

A grammar defined by means of the grammatical formalism *SUG* (*Slot Unification Grammar*) is used as input of *SUPAR*. A translator that transforms *SUG* rules into Prolog clauses has been developed. This translator will provide a Prolog program that will parse each sentence. *SUPAR* allows to carry out either a full or a partial parsing of the text, with the same parser and grammar. Here, partial parsing techniques have been used due to the unavoidable incompleteness of the grammar and the use of unrestricted texts (corpora) as inputs.

These unrestricted corpora used as input for the partial parser contain the words tagged with their grammatical categories obtained from the output of a part-of-speech (POS) tagger. The word, as it appears in the corpus, its lemma and its POS tag (with morphological information) is supplied for each word in the corpus. The corpus is split into sentences before applying the parsing.

The output of the parsing module will be the *Slot Structure* (*SS*) that stores the necessary information[2] for Natural Language Processing (NLP) problem resolution. This *SS* will be the input for the following module in which NLP problems (anaphora, extraposition, ellipsis, etc.) will be treated and solved.

In Ferrández *et al.* (1998), a partial parsing[3] strategy that provides all the necessary information for resolving anaphora is presented. This partial parsing shows that only the following constituents are necessary for anaphora resolution: co-ordinated prepositional and noun phrases, pronouns, conjunctions and

[2] The *SS* stores for each constituent the following information: constituent name (NP, PP, etc.), semantic and morphologic information, discourse marker (identifier of the entity or discourse object) and the *SS* of its subconstituents.

[3] It is important to emphasize that the system allows to carry out a full parsing of the text. In this paper, partial parsing with no semantic information is used in the evaluation of our approach.

verbs, regardless of the order in which they appear in the text. The *free words* consist of constituents that are not covered by this partial parsing (e.g. adverbs).

After applying the anaphora resolution module, a new Slot Structure (*SS'*) is obtained. In this new structure the correct antecedent (chosen from the possible candidates) for each anaphoric expression will be stored together with its morphological and semantic information. *SS'* will be the input for the Interlingua system.

3 Interlingua system (ISS)

As said before, the Interlingua system takes the *SS* of the sentence after applying the anaphora resolution module as input. This system, named *Interlingua Slot Structure* (*ISS*), generates an interlingua representation from the *SS* of the sentence.

SUPAR generates one *SS* for each sentence from the whole text and it solves intrasentential and intersentential anaphora. Then, *ISS* generates the interlingua representation of the whole text. This is one of the main advantages of *ISS* because it is possible to generate intersentential pronominal anaphora.

To begin with, *ISS* splits sentences into clauses[4]. To identify a new clause when partial parsing has been carried out, the following heuristic has been applied:

> ***H1*** *Let us assume that the beginning of a new clause has been found when a verb is parsed and a free conjunction is subsequently parsed.*

In this particular case, a *free conjunction* does not imply conjunctions that join co-ordinated noun and prepositional phrases. It refers, here, to conjunctions that are parsed in our partial parsing scheme.

Once the text has been split into clauses, the next stage is to generate the interlingua representation for clauses. We have used a complex **feature structure** for each clause. In *Figure 2* the information of the first clause of the example (1) is presented:

> *(1) The boys of the mountains were in the garden. They were catching flowers.*

[4] A clause could be defined as "*a group of words containing a verb*".

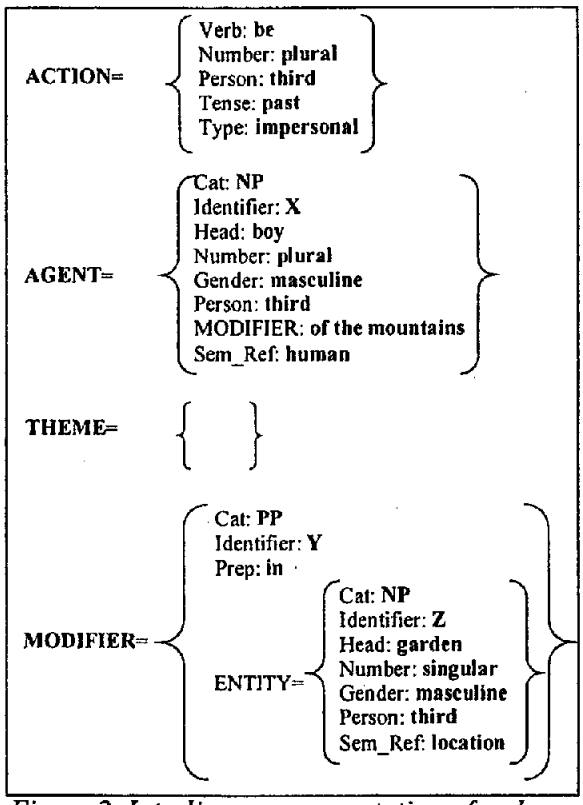

Figure 2. Interlingua representation of a clause.

As can be observed in *Figure 2*[5], the interlingua is a frame composed of semantic roles and features extracted from the *SS* of the clause. Semantic roles that have been used in this approach are the following: **ACTION**, **AGENT**, **THEME** and **MODIFIER** that correspond to verb, subject, object and prepositional phrases of the clause respectively. The notation we have used is based on the representation used in KANT interlingua. To identify these semantic roles when partial parsing has been carried out and no semantic knowledge is used, the following heuristic has been applied:

> ***H2*** *Let us assume that the NP parsed before the verb is the agent of the clause. In the same way, the NP parsed after the verb is the theme of the clause. Finally, all the PP found in the clause are its modifiers.*

[5] Only the relevant attributes of each semantic role appear in a simplified way in the picture. Additional attributes are added to the semantic roles in order to complete all the necessary information for the interlingua representation.

In *Figure 2* the following elements have been found: **ACTION**= 'were', **AGENT**= 'the boys of the mountains', **THEME**= ϕ (it has not been found any NP after the verb) and **MODIFIER**= 'in the garden'. These elements are represented by a simple feature structure. Features are represented as *attributes* with their corresponding *values*.

The semantic role **ACTION** has the following attributes: *Verb* with the value of the lemma of the verb; *Number*, *Person* and *Tense* (grammatical features) and *Type* with the type of the verb: impersonal, transitive, etc.

The semantic role **AGENT** has the following attributes: *Cat* that contains the syntactic category of the constituent; *Identifier* with the value of the discourse marker; *Head* that contains the lemma of the constituent's head; *Number*, *Gender* and *Person* contain grammatical features of the constituent; *MODIFIER* that contains all the information about the modifiers (PP) of the NP, and *Sem_Ref* that contains semantic information about the constituent's head if this information is available. The semantic role **THEME** has the same attributes as the semantic role AGENT, i.e. the difference is that THEME is the object of the clause and AGENT is the subject.

Finally, the semantic role **MODIFIER** has the following attributes: *Cat* that contains the syntactic category of the constituent; *Identifier* with the value of the discourse marker; *Prep* with the preposition of the constituent and *ENTITY*, which is the object of the PP and contains the same attributes as the **THEME**. One clause can have more than one **MODIFIER** depending on the number of PP that it has. It is important to emphasize that all this information is extracted from the *SS* of the constituents parsed in the clause.

As said before, instead of representing the clauses independently, we are interested in the interlingua representation of the whole input text. With the global representation of the input text we will be able to generate intrasentential and intersentential anaphora. Furthermore, it will be possible to solve and generate coreference chains. Thereby, the scheme of *Figure 2* is extended in order to represent all the discourse using the clauses as main units of this representation. In *Figure 3* the interlingua representation of the whole text of the example (1) can be observed.

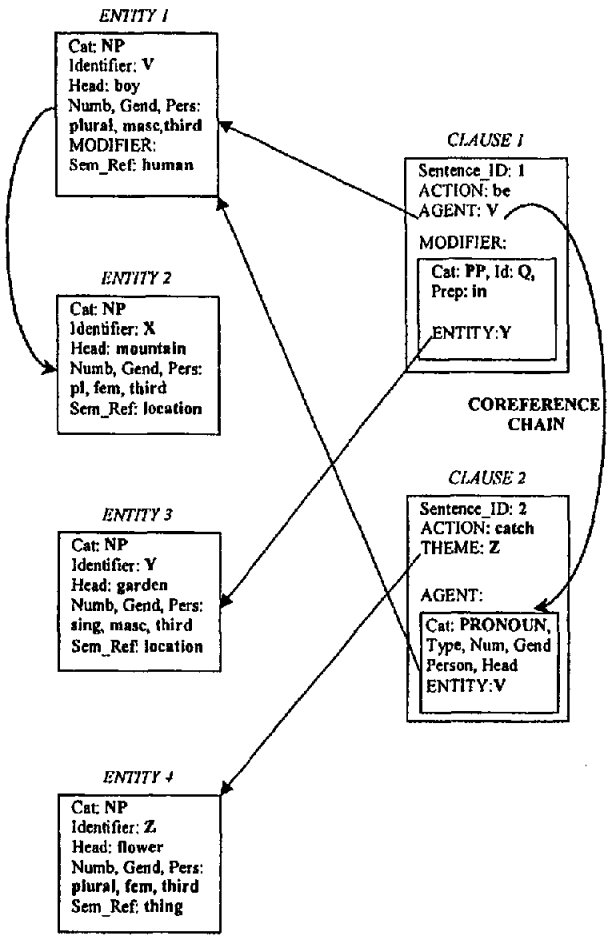

Figure 3. Interlingua representation of example (1).

On the left side of *Figure 3* the new objects or entities of the discourse are represented. These objects are named **ENTITIES** and contain the following attributes: *Cat*, *Identifier*, *Head*, *Number*, *Gender*, *Person* and *Sem_Ref*, due to they can represent an **AGENT**, a **THEME** or an object in a **MODIFIER**.

On the right side, the **CLAUSES** of the text are represented in a simplified way. They contain the semantic role **ACTION** with its attributes and the semantic roles **AGENT**, **THEME** and **MODIFIER** that have appeared in the clause. These semantic roles are linked to the **ENTITIES** that they refer to. It also contains the identifier of the sentence in which the **CLAUSE** appears (*Sentence_ID*) and the *Conjunction* that joints two or more **CLAUSES** in a sentence.

In the picture, four **ENTITIES** and two **CLAUSES** can be distinguished. The **ENTITIES** are as follows: **ENTITY 1** ('boy'), **ENTITY 2** ('mountain), **ENTITY 3** ('garden') and **ENTITY 4** ('flower'). Moreover, a relation between two **ENTITIES** (number 1 and number 2) appears in the picture due to the **ENTITY1** (NP) contains a **MODIFIER** (PP).

The **CLAUSE 1** contains: *Sentence_ID* ('1'), **ACTION** ('be'), **AGENT** ('V', the link to **ENTITY 1**) and **MODIFIER** (which is a PP and contains the link to **ENTITY 3**). The **CLAUSE 2** contains: *Sentence_ID* ('2'), **ACTION** ('catch'), **AGENT** (which is a PRONOUN and contains the link to **ENTITY 1**) and **THEME** ('Z', the link to **ENTITY 4**).

The coreference chain can be identified thanks to **AGENTS** of **CLAUSE 1** and **CLAUSE 2** ('the boys' and 'they') have their links to the same **ENTITY**. As can be seen, these links can occur between constituents of different clauses or different sentences. Then, the global system is able to generate intersentential anaphora and identify the coreference chains of the text.

4 Generation module

The Generation module takes the interlingua representation of the text as input and generates it into the target language. In this paper, we are only describing the generation of pronouns. The generation phase is split into two modules: syntactic generation and morphological generation. In the next two subsections they will be studied in detail. Although the approach presented here is multilingual, we have focused on the generation into the Spanish and English languages.

4.1 Syntactic generation

In syntactic generation the interlingua representation is converted by 'transformational rules' into an ordered surface-structure tree, with appropriate labeling of the leaves with target language grammatical functions and features. The basic task of syntactic generation is to order constituents in the correct sequence for the target language. However, the aim of this work is only the generation of pronominal anaphora into the target language, so we have only focused on the differences between the Spanish and English languages in the generation of the pronoun. These differences are what we have named *discrepancies* (a study of Spanish-English-Spanish discrepancies is showed in Peral *et al.* (1999)). In syntactic generation the following discrepancies can be found: syntactic discrepancies and Spanish elliptical zero-subject constructions.

4.1.1 Syntactic discrepancies

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 without a specific order in the sentence. In order to carry out a correct generation into English, we must firstly reorganize the Spanish sentence. Nevertheless, in the English-Spanish translation, in general, this reorganization is not necessary.

Let us see an example with the Spanish sentence

(2) A Pedro lo vi ayer.

(I saw Peter yesterday.)

In (2), the object of the verb, *A Pedro* (*to Peter*), appears before the verb (in the position of the theoretically subject) and the subject is omitted (this phenomena is usual in Spanish and it will be explained in the next subsection). The PP *A Pedro* (*to Peter*) functions as an indirect object of the verb (because it has the preposition *A (to)*). We can find out the subject since the verb is in first person and singular, so the subject would be the pronoun *Yo (I)*. 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, *Peter*, when it is moved from its theoretical place after the verb (as it occurs in this sentence).

As explained before, it is possible to identify the semantic roles (**AGENT, ACTION,** etc.) of the previous constituents in the **CLAUSE** applying a series of heuristics. Once the semantic roles of the constituents have been established, they will be stored in the interlingua representation. The generation into English will be a new clause in which the order of the constituents is the usual in English: **AGENT, ACTION, THEME** and **MODIFIERS.**

4.1.2 Elliptical zero-subject constructions (zero-pronouns)

As commented before, the Spanish language allows to omit the pronominal subject of the sentences. These omitted pronouns are usually named *zero-pronouns*. While in other languages, zero-pronouns may appear in either the subject's or the object's grammatical position, (e.g. Japanese), in Spanish texts, zero-pronouns only appear in the position of the subject. In English texts, this sort of pronoun occurs far less frequently, as the use of them are generally compulsory in the language. Nevertheless, some examples can be found: "*Ross carefully folded his trousers and Ø climbed into bed*". (The symbol Ø shows the position of the omitted pronoun). Target languages with typical elliptical (zero) constructions corresponding to source English pronouns are Italian, Thai, Chinese or Japanese.

In order to generate Spanish zero-pronouns into English, they must first be located in the text (ellipsis detection), and then resolved (anaphora resolution). At the ellipsis detection stage, information about the zero-pronoun (e.g. person, gender, and number) must first be obtained from the verb of the clause and then used to identify the antecedent of the zero-pronoun (resolution stage). The detection process depends on the knowledge about the structure of the language itself, which gives us clues to the use of each type of zero-pronoun.

The resolution of zero-pronouns has been implemented in *SUPAR*. As we may work on unrestricted texts to which partial parsing is applied, zero-pronouns must also be detected when we do not dispose of full syntactic information. Once the input text has been split into clauses after applying the heuristic ***H1***, the next problem consists of the detection of the omission of the subject from each clause.

If partial parsing techniques have been applied, we can establish the following heuristic to detect the omission of the subject from each clause:

> ***H3*** *After the sentence has been divided into clauses, a noun phrase or a pronoun is sought, for each clause, through the clause constituents on the left-hand side of the verb, unless it is imperative or impersonal. Such a noun phrase or pronoun must agree in person and number with the verb of the clause.*

Sometimes, gender information of the pronoun can be obtained when the verb is copulative. For example, in:

> *(3) Pedro$_j$ vio a* Ana$_k$ *en el parque.* Ø$_k$ *Estaba muy guapa.*
>
> *(Peter$_j$ saw* Ann$_k$ *in the park.* **She$_k$** *was very beautiful.)*

In this example, the verb *estaba (was)* is copulative, so that its subject must agree in gender and number with its object whenever the object can have either a masculine or a feminine linguistic form (*guapo: masc, guapa: fem*). We can therefore get information about its gender from the object, *guapa ("beautiful" in its feminine form)* which automatically assigns it to the feminine gender so the omitted pronoun would have to be *she* rather than *he*.

After the zero-pronoun has been detected, *SUPAR* inserts the pronoun (with its information of person, gender and number) in the position in which it has been omitted. This pronoun will be detected and resolved in the following module of anaphora resolution. After that, *ISS* generates the interlingua representation of the text.

In the example (3), two **CLAUSES** are identified. In the second **CLAUSE** the zero-pronoun is detected (third person, singular and feminine *-she-*) and solved (third person, singular and feminine *-Ann-*). So the **AGENT** of this **CLAUSE** is the PRONOUN *she* and it has a link to the **ENTITY** *Ann* (the chosen antecedent).

Now, the generation of Spanish zero-pronouns into English is easy because all the information that it is needed is located in the interlingua representation. The English pronoun's information is extracted in the following way: number and person information are obtained from the PRONOUN and gender information is obtained from the *Head* of its antecedent.

4.2 Morphological generation

In the morphological generation we mainly have to treat and solve number and gender discrepancies in the generation of pronouns.

4.2.1 Number discrepancies

This 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. For example, in English the concept *people* is plural, whereas in Spanish is singular.

> *(4) The stadium was full of* people$_i$. They$_i$ *were very angry with the referee.*
>
> *(5) El estadio estaba lleno de* gente$_i$. Ésta$_i$ *estaba muy enfadada con el árbitro.*

In (4), it can be observed that the name *people* in English has been replaced with the plural pronoun *they*, whereas in Spanish (5) the name *gente* has been replaced with the singular pronoun *ésta (it)*. Gender discrepancies also exist in the translation of other languages such as in the German-English translation.

In order to take into account number discrepancies in the generation of the pronoun into the target language a set of morphological (number) rules is constructed.

In the generation of the pronoun *They* into Spanish in the example (4), the interlingua representation has a PRONOUN ('they', third person and plural) that it is linked to the **ENTITY** ('police', plural). For the correct generation into Spanish the following morphological rule is constructed:

> *pronoun + third_person + plural + antecedent ('police') → **ésta** (pronoun, third person, feminine and singular)*

The left-hand side of the morphological rule contains the interlingua representation of the pronoun and the right-hand side contains the pronoun in the target language.

In the same way, a set of morphological rules is constructed in order to generate English pronouns. Next, an example of these rules is shown:

> *pronoun + third_person + singular + antecedent ('policía') → **they** (pronoun, third person and plural)*

4.2.2 Gender discrepancies

English has less morphological information than Spanish. With reference to plural personal pronouns, the pronoun *we* can be translated into *nosotros* (masculine) or *nosotras* (feminine), *you* into *ustedes* (masculine/feminine), *vosotros* (masculine) or *vosotras* (feminine) and *they* into *ellos* or *ellas*. Furthermore, the singular personal pronoun *it* can be translated into *él/éste* (masculine) or *ella/ésta* (feminine). For example:

> *(6)* Women$_i$ *were in the shop.* They$_i$ *were buying gifts for their husbands.*
>
> *(7)* Las mujeres$_i$ *estaban en la tienda.* Ellas$_i$ *estaban comprando regalos para sus maridos.*

In Spanish, the plural name *mujeres* (*women*) is feminine and is replaced by the personal pronoun *ellas* (plural feminine) (7), whereas in English *they* is valid for masculine as well as for feminine (6).

These discrepancies do not always mean that Spanish anaphors bear more information than English one. For example, Spanish possessive adjectives (*su casa*) do not carry gender information whereas English possessive adjectives do (*his/her house*).

We can find similar discrepancies among other languages. For example, in the French-German translation, gender is assigned arbitrarily in both languages (although in French is not as arbitrarily as in German). The English-German translation, like English-Spanish, supposes a translation from a language with neutral gender into a language that assigns gender grammatically.

As commented, it is important to emphasize that the omission of the pronominal subject is very usual in Spanish. If we want to stress the subject of a clause or distinguish between different possible subjects, we will have to write the pronominal subject. Otherwise, pronominal subject could be omitted. We are interested, however, in the correct generation of pronouns, and therefore, they will never be omitted.

Thanks to the fact that our system solves only personal pronouns in third person, we have only studied gender discrepancies in the generation of the third person pronouns. The

study has been divided into pronouns with subject role and pronouns with complement role. *a) Pronouns with subject role.* This kind of pronouns can be identified in the interlingua representation because they have the semantic role of **AGENT** in a **CLAUSE**. Their antecedents are established with the links to the **ENTITIES**.

The main problem in the pronoun generation into English consists of the generation of pronoun *it*. If we have a pronoun with the following attributes: *masculine*, *singular* and *third person* in the interlingua representation, this can be generated into the Spanish pronouns *he* or *it*. If the antecedent of the pronoun refers to a person, we will generate it into *he*. If the antecedent of the pronoun is an animal or a thing we will generate it into *it*. These characteristics of the antecedent can be obtained from the semantic information stored in its attribute *Sem_Ref*. A similar strategy is used to generate the pronouns *she* or *it*. With reference to plural personal pronouns: *masculine/feminine*, *plural* and *third person*, they are generated into the English pronoun *they*.

In *Figure 4*, the set of morphological rules to treat gender discrepancies in English generation of pronouns is shown:

```
pron + third_person + masculine + sing + antec (person) → he
pron + third_person + masculine + sing + antec (animal or thing) → it
pron + third_person + feminine + sing + antec (person) → she
pron + third_person + feminine + sing + antec (animal or thing) → it
pron + third_person + feminine/masculine + plural → they
```

Figure 4.

In Spanish generation, the main problem consists of the translation of pronoun *it* The set of morphological rules to treat this case is shown in *Figure 5*:

```
pron + third_person + sing + antec (animal with masculine gender) → él
pron + third_person + sing + antec (thing with masculine gender) → éste
pron + third_person + sing + antec (animal with feminine gender) → ella
pron + third_person + sing + antec (thing with feminine gender) → ésta
```

Figure 5.

b) Pronouns with complement role. This kind of pronouns can be identified in the interlingua representation because they have the semantic role of **THEME** or they are in a **MODIFIER** in a **CLAUSE**.

In the pronoun generation into English, the set of morphological rules of *Figure 6* is applied:

```
pron + third_person + sing + antec (person with masculine gender) → him
pron + third_person + sing + antec (person with feminine gender) → her
pron + third_person + sing + antec (animal or thing) → it
pron + third_person + plural + antec (person) → them
```

Figure 6.

In the process of generating a pronoun with the semantic role of **THEME** into Spanish, the set of morphological rules of *Figure 7* is applied:

```
pron + third_person + singular → le
pron + third_person + plural → les
```

Figure 7.

On the other hand, if the pronoun is in a **MODIFIER**, the rules of Spanish generation will be as shown in *Figure 8*:

```
pron + third_person + sing + antec (masculine gender) → él
pron + third_person + sing + antec (feminine gender) → ella
pron + third_person + plural + antec (masculine gender) → ellos
pron + third_person + plural + antec (feminine gender) → ellas
```

Figure 8.

5 Evaluation

The syntactic generation and morphological generation modules of our approach have been evaluated. To do so, one experiment for each module has been accomplished. In the first one, the generation of Spanish zero-pronouns into English, using the techniques described above in subsection 4.1.2, has been evaluated[6]. In the second one, the generation of English pronouns into Spanish ones has been evaluated. In this experiment number and gender discrepancies and their resolution, described above in section 4.2, have been taken into account.

With reference to the first experiment, our computational system has been trained with a

[6] Syntactic discrepancies has not been evaluated due to the aim of this work is only the pronominal anaphora generation into the target language, so the evaluation of the generation of the whole sentence into the target language has been omitted.

handmade corpus[7] that contains 106 zero-pronouns. With this training, we have extracted the degree of importance of the preferences that are used in the anaphora resolution module of the system. Furthermore, we have been able to check and correct the techniques used in the detection and generation of zero-pronouns into English. After that, we have carried out a blind evaluation on unrestricted texts using the set of preferences and the generation techniques learned during the training phase. In this case, partial parsing of the text with no semantic information has been used.

With regard to unrestricted texts, our system has been run on two different Spanish corpora: a) a fragment of the Spanish version of *The Blue Book* corpus (15,571 words), which contains the handbook of the International Telecommunications Union CCITT, and b) a fragment of the *Lexesp* corpus (9,746 words), which contains ten Spanish texts from different genres and authors. These texts are taken mainly from newspapers. These corpora have been POS-tagged. Having worked with different genres and disparate authors, we feel that the applicability of our proposal to other sorts of texts is assured.

To evaluate the generation of Spanish zero-pronouns into English three tasks have been accomplished: a) the evaluation of the detection of zero-pronouns, b) the evaluation of anaphora resolution and c) the evaluation of generation.

a) *Evaluating the detection of zero-pronouns.* To do this, verbs have been classified into two categories: 1) verbs whose subjects have been omitted, and 2) verbs whose subjects have not. We have obtained a success rate[8] of 88% on 1,599 classified verbs, with no significant differences seen between the corpora. We should also remark that a success rate of 98% has been obtained in the detection of verbs whose subjects were omitted, whereas only 80% was achieved for verbs whose subjects were not. This lower success rate is justified for several reasons. One important reason is the non-detection of impersonal verbs by the POS tagger. Two other reasons are the lack of semantic information and the inaccuracy of the grammar used. It is important to note that 46% of the verbs in these corpora have their subjects omitted. It shows quite clearly the importance of this phenomenon in Spanish.

b) *Evaluating anaphora resolution.* In this task, the evaluation of zero-pronoun resolution is accomplished. Of the 1,599 verbs classified in these two corpora, 734 of them have zero-pronouns. Only 228 of them[9], however, are in third person and will be anaphorically resolved. A success rate of 75% was attained for the 228 zero-pronouns. By "successful resolutions" we mean that the solutions offered by our system agree with the solutions offered by two human experts.

c) *Evaluating zero-pronoun generation.* The generation of the 228 Spanish zero-pronouns into English has been evaluated. The following results in the generation have been obtained: a success rate of 70% in *Lexesp* and a success rate of 89% in *The Blue Book*. In general (both corpora) a success rate of 75% has been achieved. The errors are mainly produced by fails in anaphora resolution and fails in the generation of pronouns *he/she/it* (some heuristics[10], which have failed sometimes, have been applied due to the used corpora do not include semantic information).

In the second experiment, we have evaluated the generation of Spanish personal pronouns with subject role into the English ones. A fragment of the English version of *The Blue Book* corpus (70,319 words) containing 165

[7] This corpus contains sentences with zero-pronouns made by different researchers of our Research Group.

[8] By "success rate", we mean the number of verbs successfully classified, divided by the total number of verbs in the text.

[9] The remaining pronouns are not in third person or they are cataphoric (the antecedent appears after the anaphor) or exophoric (the antecedent does not appear, linguistically, in the text).

[10] For instance: "all the pronouns in third person and singular whose antecedents are proper nouns have been translated into *he* (antecedent with masculine gender) or *she* (antecedent with feminine gender); otherwise they have been translated into *it*".

pronouns with subject role has been used in order to carry out a blind evaluation. A success rate of 85.41% has been achieved. The errors are mainly produced by fails in anaphora resolution and in the correct choice of the gender of the antecedent's *Head* in Spanish. With reference to the choice of the gender of the antecedent's *Head,* an electronic dictionary has been used in order to translate the original English word into the Spanish one, and subsequently, the gender is extracted from the Spanish word. Several problems have occurred when using this electronic dictionary:

1) the word to be translated does not appear in the dictionary, and therefore, a heuristic is applied to assign the gender
2) the correct sense of the English word is not chosen, and therefore, the gender could be assigned incorrectly.

Conclusion

In this paper a complete approach to solve and generate pronominal anaphora in the Spanish and English languages is presented. The approach works on unrestricted texts to which partial parsing techniques have been applied. After the parsing and solving pronominal anaphora, an interlingua representation (based on semantic roles and features) of the whole text is obtained. The representation of the whole text is one of the main advantages of our system due to several problems, that are hardly solved by the majority of MT systems, can be treated and solved. These problems are the generation of intersentential anaphora, the detection of coreference chains and the generation of Spanish zero-pronouns into English. Generation of zero-pronouns and Spanish personal pronouns has been evaluated obtaining a success rate of 75% and 85.41% respectively.

References

Appelo, L. and Landsbergen, J. (1986) The machine translation project Rose. In *Proceedings of I. International Conference on the State of the Art in Machine Translation in America, Asia and Europe, IAI-MT'86* (Saarbrücken). pp. 34-51.

Castellón, I.; Fernández, A.; Martí, M.A.; Morante, R. and Vázquez, G. (1998) An Interlingua Representation Based on the Lexico-Semantic Information. In *Proceedings of the Second AMTA SIG-IL Workshop on Interlinguas* (Philadelphia, USA, 1998).

Ferrández, A.; Palomar, M. and Moreno, L. (1998) 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, 1998). pp. 385-391.

Ferrández, A.; Palomar, M. and Moreno, L. (2000) *An empirical approach to Spanish anaphora resolution.* To appear in Machine Translation (Special Issue on anaphora resolution in Machine Translation). 2000.

Leavitt, J.R.R.; Lonsdale, D. and Franz, A. (1994) A Reasoned Interlingua for knowledge-Based Machine Translation. In *Proceedings of CSCSI-94.*

Miyoshi, H.; Ogino, T. and Sugiyama, K. (1997) EDR's Concept Classification and Description for Interlingual Representation. In *Proceedings of the First Workshop on Interlinguas* (San Diego, USA, 1997).

Peral, J.; Palomar, M. and Ferrández, A. (1999) Coreference-oriented Interlingual Slot Structure and Machine Translation. In *Proceedings of ACL Workshop on Coreference and its Applications* (College Park, Maryland, USA, 1999). pp. 69-76.

Witkam, A.P.M. (1983) *Distributed language translation: feasibility study of multilingual facility for videotex information networks.* BSO, Utrecht.