Bilingual alignment of anaphoric expressions

R. Muñoz†, R. Mitkov‡, M. Palomar†, J. Peral†, R. Evans‡, L. Moreno§
C. Orasan‡, M. Saiz-Noeda†, A. Ferrández†, C. Barbu‡,
P. Martínez-Barco† and A. Suárez†

†Grupo de Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, Spain
{rafael,mpalomar,jperal,max,antonio,patricio,armando}@dlsi.ua.es

§Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Valencia, Spain
lmoreno@dsic.upv.es

‡Computational Linguistics Group
University of Wolverhampton
Wolverhampton, UK
{r.mitkov,r.j.evans,c.orasan,c.barbu}@wlv.ac.uk

Abstract

In this paper we present an automatic mechanism for bilingual (Spanish-English) alignment of anaphoric expressions. For this purpose, two anaphora resolution systems were used. Both are based on linguistic preferences and constraints, for Spanish (SUPPAR) and for English (MARS). These systems have been independently developed and each of them is presented individually with their evaluation results. The majority of the paper presents an automatic alignment method for pronominal anaphora in Spanish and English. Once an anaphor has been solved (in both languages) this method matches anaphoric expressions and antecedents from both texts. A bitext map method has been used for the alignment with a set of bilingual texts for the evaluation. These texts have been extracted from several European Community Official documents (EUR-lex database). The alignment mechanism can be applied to different tasks related to Machine Translation such us pattern learning for translation or evaluation for automatic generation of multilingual anaphora.

1. Introduction

Tasks like pattern learning for translation or evaluation for automatic generation of multilingual anaphora in applications like Machine Translation (MT) require aligned bilingual corpora. This kind of corpus can contribute to the performance of a variety of tasks (such as the treatment of anaphora) by robust systems. Our proposal is motivated by a number of factors: most omitted pronouns in Spanish are translated into personal pronouns in English. In this case, the use of aligned corpora helps in parsing by checking for the existence of the pronoun in Spanish. Furthermore, using aligned corpora in anaphora resolution in one language facilitates the exploitation of information explicitly available in the other language. To illustrate, in Figure 4 the Spanish demonstrative pronoun “éstas” is translated into the English personal pronoun “they”. The former provides gender and number information (feminine and plural) while the latter provides only number information. So, when resolving the English pronoun “they” we can discard the NP candidate “Member States” due to the Spanish aligned pronoun “éstas” is feminine and it does not agree in gender with this candidate. We can conclude that the use of an alignment system improves the success rate of anaphora resolution process since the information associated to both pronouns, such as gender and number, are used.

2. Approaches to text alignment

Interest in automatically constructing aligned bilingual text began during the eighties, at a time when the electronic storage of large text corpora became practical for researchers. Independent efforts on the subject were being pursued in many places, including ISSCO (Catizone et al., 1989), IBM’s Thomas J. Watson research center (Brown et al., 1991), AT&T Bell Laboratories in Murray Hill (Gale and Church, 1991) and Xerox PARC (Kay and Röscheisen, 1993).

Since their inception, bilingual alignments have been shown to play a decisive role in many linguistic applications, e.g. bilingual terminology research (Dagan et al., 1993) and translator’s aids such as TransSearch, a bilingual concordancing tool (Isabelle et al., 1993), bilingual lexicography (Langlais et al., 1998; Klavans and Tzoukermann, 1995; Melamed, 1996), automatic acquisition of knowledge about translation (Brown et al., 1993), cross-language information retrieval (Nie et al., 1998), and TransCheck, a tool for automatic translation verification (Macklovitch, 1994; Macklovitch and Hannan, 1996).

A textual alignment usually represents two texts which are mutual translations in such a way that it shows the correspondences or matches between segments in the two languages. The size of the segments determines the resolu-

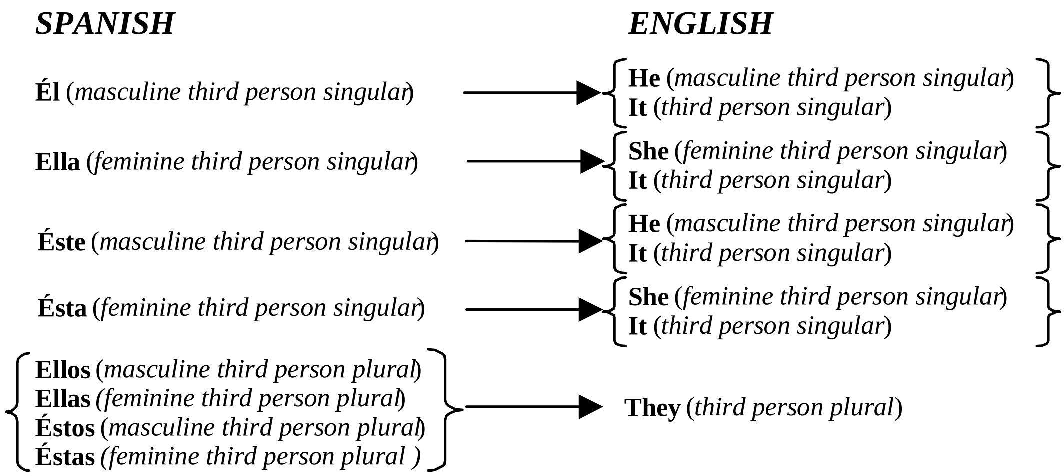

Figure 1: Gender discrepancies of pronouns between Spanish and English

tion of the alignment: paragraphs, sentences, words, etc. The result of an entire text alignment is also known as a bitext. The output of most bitext correspondence methods falls into one of two categories (as can be seen in (Simard and Plamondon, 1998)):

- An alignment is a parallel segmentation of the two texts, typically into sentences, such that the *nth* segment of the first text and the *nth* segment of the second are mutual translations.
- A bitext map is a set of pairs (x,y), where *x* and *y* refer to precise locations in the first and second texts respectively, with the intention of denoting portions of the texts that correspond to one another.

The decision to use an alignment or a bitext map usually depends on the intended application. By definition, an alignment covers the totality of the bitext. It is exhaustive and exact: for each segment of text, it pinpoints the exact segment which is its translation. Conversely, bitext maps produce approximate maps, that show that the translation of a text around point *x* is somewhere around point *y*. The maps produced are exact but not exhaustive, showing the translation of an object at position *x* is an object at position *y*.

3. Proposal of alignment

For the proposed bilingual alignment of anaphoric expressions a bitext map method has been used. In this case, the bitext map is a set of pairs (x,y), where *x* and *y* refer to precise locations of the Spanish anaphor and the English anaphor in each text respectively, with the intention of denoting anaphors of the text that correspond to one another.

In Figure 4 an example of our proposal of alignment is shown. We can see that the anaphoric expressions are aligned and will generate a pair (éstas$_1$, they$_1$). The antecedent of each one is underlined.

For the generation of the set of pairs (x,y) the following algorithm is proposed:

1. *Anaphora identification*. Pronominal anaphors of third person (for Spanish and English) and zero-pronouns[1] are identified for each anaphora resolution method.
2. *Anaphora resolution*. The anaphora resolution process is applied in order to obtain the antecedent of the anaphor in English and Spanish independently.
3. *Mechanism of characteristics*. To identify the correspondence between both anaphors, a mechanism of characteristics is applied. Due to the fact that different numbers of anaphors can appear, possibly in a different order in each text, it is necessary to perform a treatment of alignment in the framework of the paragraph. We have based this treatment on the interlingua mechanism (Peral et al., 1999) oriented to anaphora resolution and generation in Machine Translation. In this mechanism a correspondence between the English and Spanish anaphors, based on discrepancy analysis, is proposed. So, we have adapted the following four characteristics from this mechanism:

   (a) *Characteristic of number*. One problem is generated by the discrepancy between words of different languages that express the same concept. These words can be referred to by a singular pronoun in the source language and by a plural pronoun in the target language. For this case, the number discrepancy resolution presented in (Peral et al., 1999) was used.

   (b) *Characteristic of gender*. Another problem is the gender discrepancy that may exist between personal pronouns. If we identify a pronoun in a lan-

[1]Omitting the pronominal subject is very usual in Spanish. By the contrast, in English sentences, the subject is usually compulsory. This problem must be solved in the Spanish anaphora resolution method in order to obtain the corresponding English pronoun. The approach presented in (Ferr´andez and Peral, 2000) has been used for this purpose.

```
<COREF ID="valorID1"> A </COREF>
. . .
<COREF ID="valorID2" TYPE="valorTYPE" REF="valorID1"> B </COREF>
```

Figure 2: Basic SGML coreference annotation

```
Cuando los Estados miembros adopten <COREF ID="SP1"> dichas
disposiciones </COREF>, <COREF ID="SP2" TYPE="PRONOMINAL"
REF="SP1" ALIGN="ENG2"> éstas </COREF> harán referencia a la
presente Directiva o irán acompañadas de dicha referencia en su
publicación oficial.

When Member States adopt <COREF ID="ENG1"> those provisions
</COREF>, <COREF ID="ENG2" TYPE="PRONOMINAL" REF="ENG1"
ALIGN="SP2"> they </COREF> shall contain a reference to this
Directive or be accompanied by such a reference on the
occasion of their official publication.
```

Figure 3: Alignment annotation example using SGML

guage that marks its gender, it is necessary to obtain information about that pronoun's antecedent in order to identify the corresponding anaphor in the other language. In Figure 1, a correspondence of gender discrepancies between both languages is shown.

(c) *Characteristic of zero-pronouns*. As shown in (Ferrández and Peral, 2000), it is normal to omit the pronominal subject in Spanish. In these cases, we get the number and person information from the verb to obtain the corresponding English pronoun.

(d) *Characteristic of verb*. A bilingual dictionary ensures that both pronouns are linked to the same verb. If any of the verbs are not contained into the dictionary, this characteristic is skipped.

4. *Bilingual alignment of anaphoric expressions*. After the identification of the correspondence between both anaphors, we have proposed a method to annotate the bilingual alignment. This process will be performed using SGML[2].

   Gaizauskas (Gaizauskas and Humphreys, 1996) proposes a coreference annotation scheme based on SGML. According to this annotation, given an antecedent A and an anaphor B, where A and B are strings in the text, the basic coreference annotation has the form shown in Figure 2.

   The ID attribute serves to identify each string taking part in a coreference relation. The REF attribute indicates which string is coreferential with the one which it tags. The TYPE attribute serves to indicate the relationship between anaphor and antecedent.

   In our proposal, we have added a new attribute to the scheme, ALIGN, that serves to indicate the alignment between both anaphors in each language. In Figure 3, an alignment annotation example is shown.

4. Anaphora resolution systems

4.1. Anaphora resolution system in Spanish (SUPPAR)

The anaphora resolution system in Spanish (SUPPAR) has been developed to solve anaphora generated by personal (reflexive and non-reflexive), demonstrative, and zero pronouns in Spanish discourse.

SUPPAR is based on the application of a set of rules that are applied as restrictions and preferences. This approach is based on traditional restriction and preference systems such as those proposed by (Hobbs, 1978; Lappin and Leass, 1994; Kennedy and Boguraev, 1996). In these approaches, restrictions discard some candidates while preferences simply sort the remaining candidates. Our proposal differs from these since the aim of our preferences is not to sort candidates. In our system the preferences are considered similar to restrictions, except when no candidate satisfies a preference, in which case no candidate is discarded.

Several information sources are needed in order to apply this set of rules. Restrictions employ morphological, lexical and syntactic information while preferences use syntactic information. In order to obtain this information the POS Tagger by (Pla et al., 2000) and the SUPP partial parser by (Palomar et al., 1999) are used.

Over candidates identified in such analysed text, the system applies a set of restrictions based on *morphological agreement* (person, gender and number features) and *syntactic conditions* on NP-pronoun non coreference (C-command and Minimal Governing Category restrictions formulated in (Reinhart, 1983)) adapted for the syntactic information provided by the SUPP partial parser and for Spanish (Ferrández et al., 1998).

The set of preferences depends on the kind of pronoun

[2] SGML (Standard Generalized Markup Language) is a language of generation of tagging normalized by ISO8879:1986 (Goldfarb, 1990).

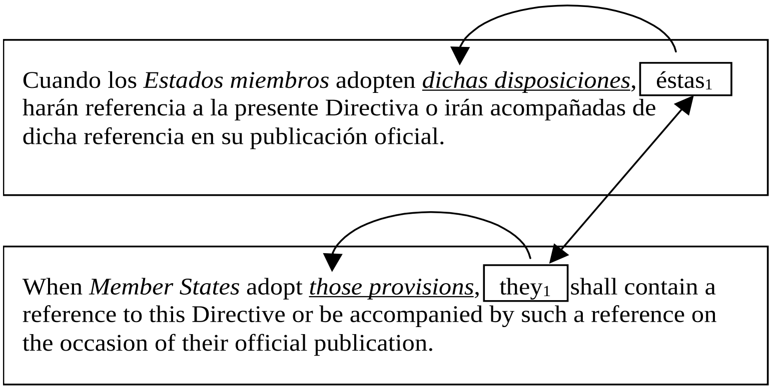

Figure 4: Example of Alignment of anaphoric expressions

to be resolved. Different sets of preferences have been defined to treat ommited, personal, or demostrative pronouns. All of them are based on linguistic information provided by the SUPP partial parser, such as the position of the candidate in the sentence relative to the verb (that is, after or before the verb), repetition of the candidate in the text, appearance of the candidate with the same verb as the anaphor, and the proximity of the candidate with respect to the anaphor.

For the proper evaluation of SUPPAR, two different corpora were selected: a technical manual on telecommunications (the Spanish version of The Blue Book[3]) that has 375 pronouns, and a collection of different kinds of text, such as journal papers and narrative texts (LEXESP[4]) with 1302 pronouns. In these corpora, a success rate of 76.8% was attained for anaphora resolution. By "success rate", we mean the number of pronouns successfully resolved, divided by the total number of resolved pronouns. The total number of resolved pronouns was 1,677. These were personal, demonstrative, reflexive and omitted pronouns. All of them were in the third person, with a noun phrase that appeared before the anaphor as their antecedent.

4.2. MARS: The English language pronominal anaphora resolution system

Pronouns in English texts were processed using an updated version of MARS, a system based on the robust pronoun resolution algorithm of (Mitkov, 1998). A detailed description of the system appears in (Mitkov et al., 2002) To recap briefly, the system runs through five phases.

In phase 1, the text to be processed is parsed syntactically, using Conexor's FDG Parser (Tapanainen and Järvinen, 1997) which returns the parts of speech, morphological lemmas, syntactic roles, grammatical number, and partial dependency relations between tokens in the text.

In phase 2, third person pronouns are identified and pleonastic instances of *it* are filtered using the machine learning method described in (Evans, 2001). MARS does not attempt to resolve demonstratives or reflexives but does attempt to resolve personal and possessive pronouns.

In phase 3, for each pronoun encountered, potential antecedents (*candidates*), are extracted from the current and preceding two sentences and the section heading. Candidates are expected to agree with their pronoun in terms of number and gender. In an improvement to the original system, intrasentential candidates must also obey syntactic constraints, implemented by exploiting the functional dependencies between NPs, verbs, and pronouns as well as the syntactic function of NPs. Similar constraints were usefully employed by (Kennedy and Boguraev, 1996) in their anaphora resolution system.

In phase 4, preferential and impeding indicators are applied to the sets of candidates. The indicators apply a numerical score to each candidate, reflecting the confidence that the candidate is the antecedent of the current pronoun. The twelve originally stated indicators were described in (Mitkov, 1998) and two additions were described in (Orasan et al., 2000). Below, we describe the new versions of four indicators and the addition of a new preferential indicator.

Frequent Candidates is the new indicator. It awards a boosting score (+1) to the three NPs that occur most frequently in the candidate sets of all pronouns in the text.

Collocation Pattern Preference now awards a boosting score of +2 to candidates matching NPs in the entire text that appear as arguments of verbs or prepositions morphologically related to verbs or prepositions that have the current pronoun as an argument.

Term Preference boosts NPs representing significant terms in the domain of the document. In the current implementation, it is the ten NPs that appear with greatest frequency in the text that are awarded the boosting score of +1.

Penalise pronoun has been renamed *boost pronoun* and applies a score of +1 to pronoun candidates. We have found that pronouns are frequently linked to distant nominal an-

[3]CRATER (1994) Corpus Resources and Terminology Extraction Project. Project supported by the European Community Commission (DG-XIII). Computational Linguistics Laboratory, Faculty Philosophy and Fine Arts, Universidad Aut´onoma de Madrid, Spain.

[4]This corpus belongs to the project of the same name carried out by the Psychology Department of the Universidad de Oviedo and developed by the Computational Linguistics Group of the Universidad de Barcelona, with the collaboration of the Language Processing Group of the Universidad Polit´ecnica de Cataluña, Spain.

tecedents via chains of pronominal mentions.

First NPs has been renamed *obliqueness*. Following works such as (Walker and Prince, 1995) where grammatical function is used as an indicator of discourse salience, it now awards subject NPs a score of +2, objects a score of +1, indirect objects no bonus, and NPs for which the FDG parser is unable to identify a function a penalizing score of -1.

Finally, in phase 5, the candidate with the highest aggregate score is selected as the antecedent of the pronoun. Ties are resolved by selecting the most recent highest scoring candidate.

MARS was evaluated with respect to eight texts from the domains of computer hardware and software manuals. The texts contained 247,401 words with 2263 anaphoric pronouns. Applied over this corpus, MARS obtained a success rate of 59.35%, with scores varying between 82.67% and 51.59% over different texts.

5. Experimental work

5.1. Description of the corpora

The data for the evaluation was taken from the EUR-Lex database (European Union law), available on http://europa.eu.int/eur-lex. We have chosen a corpus of four bilingual documents for this experiment. These corpora are official documents related to official journals, treaties, legislation, case-law, etc.

Of the four corpora, one was selected for training and the remaining three were reserved for the final evaluation.

The training corpus contains 11.135 words and 252 pronouns[5] for Spanish, and 9.801 words and 71[6] pronouns for English.

The testing corpus contains 50.871 words and 1.109 pronouns for Spanish and 43.307 words and 276 pronouns for English.

5.2. Training and evaluation phases

In this phase, the system was trained in order to obtain the characteristics of the mechanism. The training consisted of two stages: a) identification of the correspondences between pairs of anaphors (English and Spanish), b) Extraction of the main characteristics that facilitate identification of these correspondences.

Following the training phase we drew several conclusions.

Firstly, the sentence alignment was quite different between both languages, in most cases the aligned sentence numbers did not match. Therefore, we can only align anaphors by using the *mechanism of characteristics* (Section 4.2). Moreover, on examination, there were fewer words and sentences in the English corpora than the Spanish.

Secondly, there are some cases in which the kind of anaphor is not the same in both languages, for example:

.. Establecer directrices de la Comisión para su aplicación ..

*.. Establish Comission guidelines for applying **it** ..*

[5]Including omitted pronouns.

[6]Including pleonastic *it*.

In which the Spanish anaphor is a possesive determiner whereas the English anaphor is a personal pronoun.

Thirdly, the alignment has been found very useful in detecting non-anaphoric pronouns, e.g. the Spanish reflexive pronoun *se* that does not refer to any noun phrase:

.. Los responsables políticos **se** enfrentan constantemente al dilema de ..

.. Thus, decision-makers are constantly faced with the dilemma of ..

Or those zero-pronouns in which the subject of the sentence appears after the verb, as in the following example:

.. cómo $Ø_i$ debe ser la aplicación de este principio$_i$ ..

.. on how this principle$_i$ should be applied ..

During the training phase, we obtained the characteristics proposed in the algorithm described in Section 4.2.

In the training corpus a success rate of 91.3% was obtained.

Finally, the characteristics obtained in the training phase were applied in the evaluation phase. A success rate of 89.7% was obtained.

6. Conclusions

In this paper an automatic mechanism for bilingual (Spanish-English) alignment of anaphoric expressions has been presented. For this alignment a bitext map method was used. The mechanism has been evaluated on a set of bilingual texts; a success rate around 90% was obtained over the test corpus.

In the training phase we have observed different problems with the alignment of anaphoric expressions including: discrepancies in the number of sentences, the use of different pronouns, omitted pronouns in Spanish, and problems with the translation of pronouns.

7. References

P.F. Brown, J.C. Lai, and R.L. Mercer. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley CA.

P.F. Brown, S.A.D. Pietra, , V.J.D. Pietra, and R. Mercer. 1993. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2).

R. Catizone, G. Russell, and S. Warwick. 1989. Deriving Translation Data from Bilingual Texts. In *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, MI.

I. Dagan, K.W. Church, and W.A. Gale. 1993. Robust Bilingual Word Alignment for Machine Aided Translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus OH.

R. Evans. 2001. Applying machine learnong toward an automatic classification of *it*. *Journal of Literary and Linguistic Computing*, 16(1):45 – 57.

A. Ferrández and J. Peral. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics*, pages 166–172, Hong Kong, October.

A. Ferrández, M. Palomar, and L. Moreno. 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)*, pages 385–391, Montreal (Canada), August.

R. Gaizauskas and K. Humphreys. 1996. Quantitative Evaluation of Coreference Algorithms in an Information Extraction System. In *Proceedings of the Discurse Anaphora and Anaphor Resolution Colloquium (DAARC)*, Lancaster.

W.A. Gale and K.W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley CA.

C.F. Goldfarb. 1990. *The SGML Handbook*. Oxford University Press.

J. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

P. Isabelle, M. Dymetman, G. Foster, J.M. Jutras, E. Macklovitch, F. Perrault, X. Ren, and M. Simard. 1993. Translation Analysis and Translation Automation. In *Proceeding of TMI'93*, Kyoto, Japan.

M. Kay and M. Röscheisen. 1993. Text-translation Alignment. *Computational Linguistics*, 19(1).

C. Kennedy and B. Boguraev. 1996. Anaphora for Everyone: Pronominal Anaphora resolution without a Parser. In *Proceedings of 16th International Conference on Computational Linguistics (COLING'96)*, pages 113–118, Copenhagen, Denmark.

J. Klavans and E. Tzoukermann. 1995. Combining Corpus and Machine-readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, 10(3).

P. Langlais, M. Simard, and J. Véronis. 1998. Methods and Practical Issues in Evaluating Alignment Techniques. 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.

S. Lappin and H.J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

E. Macklovitch and M.L. Hannan. 1996. Line'Em Up: Advances In Alignment Technology And Their Impact on Translation Support Tools. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*, Montréal, Québec.

E. Macklovitch. 1994. Using Bi-textual Alignment for Traslation Validation: the TransChechk System. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-94)*, Columbia, Maryland.

I.D. Melamed. 1996. A Geometric Approach to Mapping Bitext Correspondence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA.

R. Mitkov, R. Evans, and C. Orasan. 2002. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, pages 168–173.

R. Mitkov. 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)*, Montreal (Canada), August.

J.Y Nie, P. Isabelle, P. Plamondon, and G. Foster. 1998. Using a Probabilistic Translation Model for Cross-language Information Retrieval. In *Proceedings of the 6th ACL Workshop on Very Large Corpora*, Montréal, Canada.

C. Orasan, R. Evans, and R. Mitkov. 2000. Enhancing preference-based anaphora resolution with genetic algorithms. In *Proceedings of Natural Language Processing - NLP 2000*, pages 185 – 195. Springer Verlag.

M. Palomar, A. Ferrández, L. Moreno, M. Saiz-Noeda, R. Mu¯noz, P. Martínez-Barco, J. Peral, and B. Navarro. 1999. A Robust Partial Parsing Strategy based on the Slot Unification Grammars. In *Proceedings of the Sixth Conference on Natural Language Processing (TALN'99)*, pages 263–272, Corsica, France, July.

J. Peral, M. Palomar, and A. Ferrández. 1999. Coreference-oriented Interlingual Slot Structure and Machine Translation. In *Proceedings of ACL workshop Coreference and its Applications*, pages 69–76, Maryland (USA), June.

F. Pla, A. Molina, and N. Prieto. 2000. Tagging and Chunking with Bigrams. In *Proceedings of 18th International Conference on Computational Linguistics (COLING'2000)*, Saarbrucken, Germany.

T. Reinhart. 1983. *Anaphora and Semantic Interpretation*. Croom Helm, London and Sydney.

M. Simard and P. Plamondon. 1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation*, 13(1):59–80.

P. Tapanainen and T. Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing*, pages 64 – 71.

M. Walker and E. Prince. 1995. A bilateral approach to givenness: A hearer-status algorithm and a centering algorithm. In J. Gundel. and T. Fretheim, editors, *Reference Accessibility*.