Exploiting semantic information for manual anaphoric annotation in Cast3LB corpus

Borja Navarro, Rubén Izquierdo, Maximiliano Saiz-Noeda
Departmento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Ap. Correo 99, E-03080
Alicante, Spain
{borja,ruben,max}@dlsi.ua.es

Abstract

This paper presents the discourse annotation followed in Cast3LB, a Spanish corpus annotated with several information sources (morphological, syntactic, semantic and coreferential) at syntactic, semantic and discourse level. 3LB annotation scheme has been developed for three languages (Spanish, Catalan and Basque). Human annotators have used a set of tagging techniques and protocols. Several tools have provided them with a friendly annotation scheme. At discourse level, anaphoric and coreference expressions are annotated. One of the most interesting contributions to this annotation scenario is the enriched anaphora resolution module that is based on the previously defined semantic annotation phase to expand the discourse information and use it to suggest the correct antecedent of an anaphora to the annotator. This paper describes the relevance of the semantic tags in the discourse annotation in Spanish corpus Cast3LB and shows both levels and tools in the mentioned discourse annotation scheme.

1 Introduction

Cast3LB corpus is annotated (Navarro et al., 2003) at three linguistic levels: sentence level (syntactic), lexical level (semantic) and discourse level. At discourse level, it is annotated with anaphoric and coreferential information. In order to improve the time-consuming and tedious task of the manually annotation, a semiautomatic and interactive process is followed: first, an anaphora resolution system selects each anaphora and its antecedent from a list of candidates; then, the human annotator decides wether or not accept the suggestion.

With this approach, the correctness of the anaphora resolution system is a key factor in the quest of an efficient annotation process. For this reason, we use the linguistic information of the previous annotation tasks (morphological, syntactic and, mainly, semantic information) to improve the anaphora resolution system. In this paper we will focus on the use of semantic information in the anaphora and coreferential manual annotation task.

Next section presents the project overview and the three annotation levels. Following sections present the semantic annotation and the way it serves to the discourse annotations. Last section presents annotation tools used to annotate the corpus at semantic and coreferential level.

2 Cast3LB corpus: annotation project overview

Cast3Lb project is part of the general project 3LB[1]. The main objective of this general project is to develop three corpora annotated with syntactic, semantic and pragmatic/coreferential information: one for Catalan (Cat3LB), one for Basque (Eus3LB) and one for Spanish (Cast3LB).

The Spanish corpus Cast3LB is a part of the CLIC-TALP corpus, which is made up of 100.000 words from the LexEsp corpus (Sebastián et al., 2000) plus 25.000 words coming from the EFE Spanish Corpus, given by the Agencia EFE (the official news agency) for research purposes. The EFE corpus fragments are comparable among the languages of the general project (Catalan, Basque and Spanish).

We have selected this corpus because it contains a large variety of Spanish texts (newspapers, novels, scientific papers…), both from Spain and South-America, so it is a good representation of the current state of the Spanish language. Moreover, the automatic morphological annotation of this corpus has been manually checked (Civit, 2003).

The spirit of the annotation scheme is to build a flexible system portable to different romance languages and to potential new cases that might appear, but consistent with all annotation levels and annotation data.

At the syntactic level we follow the constituency annotation scheme. Main principles of syntactic annotation are the following (Civit et al., 2003): a) only the explicit elements are annotated (except for elliptical subjects); b) we do not alter the surface word order of the elements; c) we do not follow any specific theoretical framework; d) we do not take into account the verbal phrase, rather, the main constituents of the sentence become the daughters of the root node; e) this syntactic information is enriched by the functional information of the main phrases, but we have not taken into account the possibility of double functions.

At the semantic level, we annotate the sense of the nouns, verbs and some adjectives, following an all words approach. The specific sense (or senses) of each one is assigned by means of the EuroWordNet offset number (Vossen, 1998). Also, due to some words are not available in EuroWordNet or do not have the suitable sense,

[1] Project partially funded by Spanish Government FIT-150-500-2002-244.

we have created two new tags to mark this circumstance.

At the discourse level, we mark the coreference of nominal phrases and some elliptical elements. The coreference expressions taken into account are personal pronouns, clitics, elliptical subjects and some elliptical adjectives. The definite descriptions are not marked. The possible antecedents considered are the nominal phrases or other coreferential expressions.

3 Semantic annotation

As we said before, main objective of Cast3LB project at semantic level is to develop an "all words" corpus with the specific sense (or senses) of nouns, verbs and adjectives.

Our proposal is based on the SemCor corpus (Miller, 1990). This corpus is formed by a portion of the Brown corpus and the novel *The Red Badge of Courage*. Altogether, it is formed by approximately 250.000 words, where nouns, verbs, adjectives and adverbs have been manually annotated with WordNet senses (Miller, 1990). Another corpus with WordNet-based semantic annotation is the DSO corpus (Ng and Lee, 1996). In this corpus, the most frequent English ambiguous nouns and verbs had been annotated with the correct sense (121 nouns and 70 verbs). The corpus is formed by 192.800 sentences from the Brown Corpus and the Wall Street Journal, and it has also been manually annotated. Finally, the SENSEVAL forum has developed a few sense annotated corpora for the evaluation of Word Sense Disambiguation systems (Kilgarriff and Palmer, 2000), some of which also use WordNet as a lexical resource.

We have decided to use Spanish WordNet for several reasons. First of all, Spanish WordNet is, up to now, the more commonly used lexical resource in Word Sense Disambiguation tasks. Secondly, it is one of the most complete lexical resources currently available for Spanish. Finally, as part of EuroWordNet, the lexical structure of Spanish and the lexical structure of Catalan and Basque are related. Therefore, the annotated senses of the three corpora of 3LB project can also be related.

The tag used to mark a word sense is its offset number, that is, its identification number in EuroWordNet's InterLingua Index. The corpus has 42291 lexical words, where 20461 are nouns, 13471 are verbs and 8543 are adjectives.

On other hand, not all nouns, verbs, adjectives and adverbs are annotated, due to EuroWordNet does not contain them. Possible lacks in this sense are (i) the synset, (ii) the word, (iii) the synset and the word, and (iv) the link between the synset and the word.

In order to deal with these cases we have defined two more tags in EuroWordNet:

- C1S: the word is found, but not its correct sense (due to a sense lack, or because there is no link between the word and the synset).
- C2S: the word is not found (because it is not there, or because both the word and the synset are missing).

It is possible to distinguish two methods for semantically annotate a corpus. The first one is linear (or "textual") method (Kilgarriff, 1998), where the human annotator marks the sentences token by token up to the end of the corpus. In this strategy the annotator must read and analyze the sense of each word every time it appears in the corpus. The second annotation method is transversal (or "lexical") (Kilgarriff, 1998), where he/she annotates word-type by word-type, all the occurrences of each word in the corpus one by one. With this method, the annotator must read and analyze all the senses of a word only once.

We have followed in Cast3LB the transversal process. The main advantage of this method is that we can focus our attention on the sense structure of one word and deal with its specific semantic problems: its main sense or senses, its specific senses. . . . Then we check the context of the single word each time it appears and select the corresponding sense. Through this approach, semantic features of each word is taken into consideration only once, and the whole corpus achieves greater consistency. Through the linear process, however, the annotator must remember the sense structure of each word and their specific problems each time the word appears in the corpus, making the annotation process much more complex, and increasing the possibilities of low consistency and disagreement between the annotators.

Nevertheless, the transversal method finds its disadvantage in the annotation of large corpus, because no fragment of the corpus is available until the whole corpus is completed. To avoid this, we have selected a fragment of the whole corpus and annotated it by means of the linear process.

Everybody agrees that semantic annotation is a tedious and difficult task. From a general point of view, the main problem in the semantic annotation is the subjectivity of the human annotator when it comes to the selection of the correct sense, because there are usually more than one sense for a word, and, due to the WorNet's granularity, more than one could be correct for a given word. Another important problem in the semantic annotation is the poor agreement between different annotators, due to the ambiguity and/or vagueness of many words.

In order to overcome these problems, the annotation process has been carried out in two steps. In the first step, a subset of ambiguous words have been annotated twice by two annotators. With this double annotation we have developed a disagreement typology and an annotation handbook, where all the possible causes of ambiguity have been described and common solutions have been adopted for the rest of cases. In the second step the remaining corpus is annotated following the criteria adopted in the annotation handbook.

Our final aim is to obtain useful resources for Word Sense Disambiguation (WSD) systems in Spanish. This semantically annotated corpus will be used as a training corpus for the development of unsupervised systems and as a reference in general evaluation tasks. At the end of the project, we will have a large amount of words with an unambiguous sense tag in a real context.

As well as this final application, we exploit this semantic information in the anaphoric annotation task. In (Saiz-Noeda, 2002), how to apply semantic information in anaphora resolution systems is showed and evaluated. We take this proposal, but applied to manual anaphora annotation.

Due to the corpus has been annotated with syntactic information, and the sense of each word is marked with the offset number of EuroWordNet, it is possible to extract semantic features of each verb and noun through the ontological concepts of the EuroWordNet's Top Ontology. Furthermore, the corpus has been annotated with syntactic roles, so it is possible to extract syntactic patterns formed by the verb and its main complements: subject-verb, verb-direct objects, verb-indirect objects.

As we will show bellow, these patterns are useful in order to select the specific antecedent of an anaphora, according to semantic compatibility criteria between the antecedent and the verb of the sentence where the anaphora appears.

4 Discourse annotation: anaphora and coreference

At discourse level, our objective is to annotate the anaphora and the coreference, in order to develop useful resources for anaphora resolution systems.

We agreed to annotate the anaphoric elements and their antecedents. These anaphoric elements are the anaphoric ellipsis, the pronominal anaphora and the coreferential chains.

Specifically, in each one, we mark:

- Anaphoric ellipsis:
  - The elliptical subject, made explicit in the syntactic annotation step. Being a noun phrase, it could also be an antecedent too.

    Unlike English, where it is possible an expletive pronoun as subject, in Spanish it is very common an elliptical nominal phrase as subject of the sentence. This is why we have decide to include this kind of anaphora in the annotation process.
  - Elliptical head of nominal phrases with an adjective complement. In English, this construction is the "one anaphora". In Spanish, however, the anaphoric construction is made up by an elliptical head noun and an adjective complement.
- Anaphora: Two kinds of pronouns:
  - The tonic personal pronouns in the third person. They can appear in subject function or in object function.
  - The atonic pronouns, specifically the clitic pronouns that appear in the subcategorization frame of the main verb.
- Finally, there are sets of anaphoric and elliptical units that corefer to the same entity. These units form coreferential chains. They must be marked in order to show the cohesion and coherence of the text. They are annotated by means of the identification of the same antecedent.

We do not annotate the definite descriptions. They consist of nominal phrases that can refer (or not) to an antecedent. We do not mark them because they outline specific problems that make this task very difficult: firstly, there are not clear criteria that allow us to distinguish between coreferential and not coreferential nominal phrases; secondly, there are not a clear typology for definite descriptions; and finally, there are not a clear typology of relationships between the definite description and their antecedents. These problems could further increase the time-consuming in the annotation process and widen the gap of disagreement between the human annotators.

This proposal of annotation scheme is based on the one used in the MUC (Message Understanding Conference) (Hirschman, 1997) as well as in the works of Gaizauskas (Gaizauskas and Humphreys, 1996) and Mitkov (Mitkov et al., 2002): this is the mostly used scheme in coreferential annotation (Mitkov, 2002).

In the anaphoric annotation, two linguistic elements must be marked: the anaphoric expression and its antecedent. In the antecedent we annotate the following information:

- A reference tag that shows the presence of an antecedent ("REF"),
- An identification number ("ID"),
- The minimum continuous substring that could be considerer correct ("MIN").

In the coreferential expression, we annotate:

- The presence of a coreferential expression ("COREF"),
- An identification number ("ID"),
- The type of anaphoric expression: elliptical subject, elliptical head of noun phrase, tonic pronoun or atonic pronoun ("TYPE"),
- The antecedent, through its identification number ("REF"),
- Finally, a status tag where the annotators shows their confidence in the annotation ("STATUS").

As previously mentioned in this paper, the main problem in the anaphoric annotation is the low agreement between human annotators. There is usually less agreement in anaphoric annotation than in syntactic annotation ((Mitkov, 2002), 141). In order to reduce this low agreement, we annotate only the clearest type of anaphoric units (pronouns, elliptical subjects and elliptical nominal heads), and we introduce the lowest necessary information. Moreover, with the tag "STATUS", the human annotator can show his confidence in the anaphoric unit and the antecedent marked. However, at the moment, as occurs in the semantic annotation, we do not have enough data on the agreement between annotators.

4.1 Manual annotation with an Enriched Anaphora Resolution System

As we said before, we follow a manual anaphora annotation with the help of a Enriched Anaphora Resolution System: our idea is to check the automatic annotation of the anaphora resolution system and to correct mistakes in the annotation process.

In manual anaphora and coreferential annotation, the human annotator first locates a possible anaphora, and then must read back the text until the antecedent appears. With an anaphora resolution system it is possible to automatize this process: the system selects possible anaphoric elements, their possible antecedents, and decides the main candidate. The human annotator must only check the suggestion. The process is more useful because the most tedious task (to select a possible anaphora, to read back looking for the antecedent, etc.) is made up by the system. When the human annotator checks the solution, he does not read back for antecedents, he goes directly to the possible antecedents.

However, the anaphora resolution system must be very accurate. In order to automatically specify the antecedent of an anaphora and ensure the correctness of the system, we use all the linguistic information previously annotated in the corpus: morphological, syntactic and semantic. In this knowledge-based anaphora resolution system, the linguistic information is used through a set of restrictions and preferences. Following this strategy, the system rejects possible antecedents until only one is selected. The key point is the linguistic information used in restrictions and preferences.

We have developed a semantically enriched anaphora resolution system in order to aid the discourse annotation level. EuroWordNet synsets are the base of the semantic information added to the resolution process. The fact of counting with a semantically annotated corpus such as Cast3Lb facilitates the use of the anaphora resolution method, based on a natural way of understanding the human process for anaphora resolution.

The specific use of semantic information is related to the sematic compatibility between the possible antecedent (a noun) and the verb of the sentence in which the anaphoric pronoun appears. Due to the pronoun replaces a lexical word (the antecedent), the semantic information of the antecedent must be compatible with the semantic restrictions of the verb. In other words, the anaphoric expression takes the semantic features of the antecedent, so they must be compatible with the semantic restrictions of the verb.

In this way, verbs like “eat” or “drink” will be specially compatible with animal subjects and eatable and drinkable objects than others.

In our case, the semantic features of the lexical words have been extracted form the ontological concepts of EuroWorNet, that is, the Top Ontology. All the synsets in EuroWordnet are semantically described through a set of base concepts (the more general concepts). In the EuroWorNet’s Top Ontology, these base concepts are classified in the three orders of Lyons (Lyons, 1977), according to basic semantic distinctions. So through the top ontology, all the synsets of EuroWordNet are semantically described with concepts like “human”, “animal”, “artifact”, etc. With this, we have extracted subject-verb, verb-direct object and/or verb-indirect object semantic patterns.

From this semantic patters, rules about the semantic compatibility between nouns and verbs have been extracted. These rules are applied to the anaphora resolution as preferences. Based on the patterns, the system calculates the compatibility between the verb of the sentence in which the anaphora appears and the antecedent. So the possible antecedents with low compatibility are rejected, and the antecedents with high compatibility are selected. These semantic preferences, plus the syntactic and morphological restrictions and preferences, are used to select the correct antecedent of the anaphora.

Furthermore, semantic information is also used in some rules. There are two kind of rules:

- “NO” rules: NO(v#sense,c,r) defines the incompatibility between the verb v (and it sense) and any name which contains ’c’ in its ontological concept list, being ’r’ the syntactic function that relates them.
- “MUST” rules: MUST(v#sense,c,r) defines the incompatibility between the verb v (and its sense) and all the names that don’t contain ’c’ in their ontological concept list, being ’r’ the syntactic function that relates them.

At the final annotation step, the annotator checks if the antecedent selected is the correct one or not, and, in each case, confirms the annotation or corrects it.

5 Tools

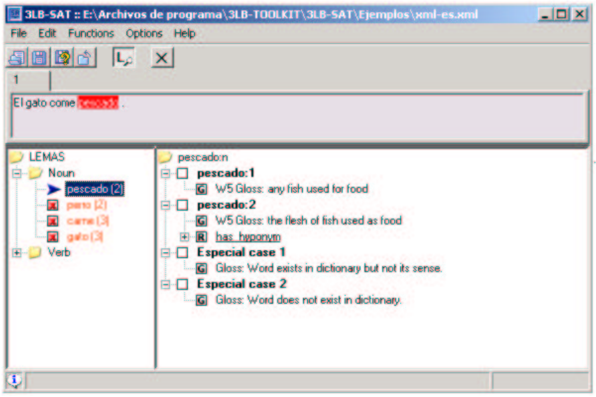
5.1 3LB-SAT

3LB-SAT (Semantic Annotation Tool) is a tool for the semantic tagging of multilingual corpora. Main features of this tool are:

- it is word-oriented,
- it allows different format for input corpus; basically, the main formats used in corpus annotation: treebank format (TBF) and XML format;
- it uses EuroWordNet as a lexical resource.

For the XML format a DTD has been defined, that allows to describe the information structure in each file of the corpus.

In the annotation process, monosemic words are automatically annotated. So, 3LB-SAT is used to annotated only the polysemic words. When a file is loaded, all lemmas of the file are shown (Figure 1). The tool uses different colors to indicate the state of the annotation process: (i) no occurrence of the lemma in the file has been annotated, (ii) some occurrences of the lemma in the file have been annotated, or (iii) all the occurrences have been annotated. When the annotator selects a lemma, all its occurrences are shown. The selection of one of them shows all possible senses, and the annotator chooses the correct one for this specific context.


3LB-SAT :: E:\Archivos de programa\3LB-TOOLKIT\3LB-SAT\Ejemplos\xml-es.xml
File
Edit
Functions
Options
Help
1

El gato come pescado .

- LEMAS
  - Noun
    - pescado (2)
    - pierto (2)
    - carne (3)
    - gato (3)
  - Verb

- pescado:n
  - □ **pescado:1**
    - G W5 Gloss: any fish used for food
  - □ **pescado:2**
    - G W5 Gloss: the flesh of fish used as food
    - R has_hyponym
  - □ **Especial case 1**
    - G Gloss: Word exists in dictionary but not its sense.
  - □ **Especial case 2**
    - G Gloss: Word does not exist in dictionary.

Figure 1: 3LB-SAT semantic annotation tool.

5.2 3LB-RAT

3LB-RAT (Reference Annotation Tool) is a tool developed in 3LB project for the annotation and supervision of anaphora and coreferences at discourse level.

The tool provides the annotator with two working ways: manual and semiautomatic. In the first one, the tool locates and shows all possible anaphoric and coreference elements and their possible antecedents. The annotator chooses one of these possible antecedents and indicates the certainty degree on this selection (standby, certain or uncertain).

There are some exceptional cases that the tool always offers:

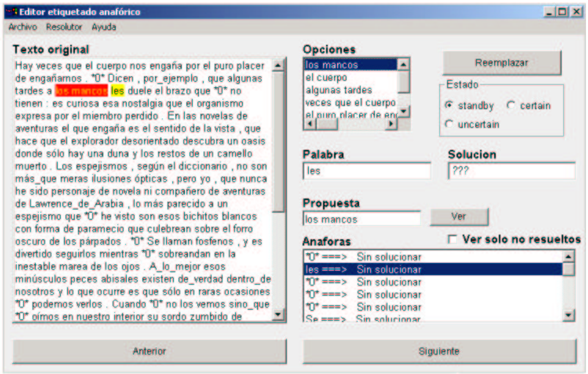
- cases of cataphora,
- possible syntactic mistakes (that will be used to review and to correct the syntactic annotation),
- the possibility of a non-located antecedent,
- the possibility that an antecedent doesn't appear explicitly in the text,
- the possibility of non-anaphora, that is, the system has not correctly located an anaphoric expression.

In the semiautomatic way, the tool solves each coreference by means of the enriched resolution anaphora method previously explained. So the system proposes and shows the most suitable candidate to the annotator. The annotator can choose the solution that the resolution method offers in all cases, or choose another solution (manually).

3LB-RAT has been developed in Python language, which guarantees the portability to any Windows or Unix platform. It deals with XML files: it is designed to work and to understand the format used by the 3LB-SAT tool, but it is able to accept any other XML specification.

As we said before, the tool uses syntactic, morphologic and semantic information for the specification of an anaphora and its antecedent. The semantic information used by the tool is limited to ontology concepts and synonymous. From the semantically annotated text, three tables are created, one for each syntactic function: subject, direct object and indirect object. In these tables the appearance frequency of nouns with verbs (with their correct senses) is stored. These tables are the base to construct the semantic compatibility patterns, which indicate the compatibility between the ontological concept related with the possible antecedent and the verb of the sentence where the anaphoric expression appears. In order to calculate this information, the occurrence frequency and the conceptual generality degree in the ontology are considered. In this case, a higher punctuation is given to the most concrete concepts. For example, "Human" concept gives us further information than "Natural" concept. These patterns are used in the semantic preferences application. For a specific candidate, its semantic compatibility is calculated from the compatible ontological concepts on the patterns. The candidates with greater compatibility are preferred.

When the annotator selects a XML file to open, the possible anaphoric elements of the text and their candidates are located, and each anaphora is solved. The system shows two lists (Figure 2): the lower list shows each anaphora located and its solution. When the annotator selects one of these elements, in the upper box appears the possible candidates list besides the solution suggested by the system. At the same time, in the plain text, the anaphora and the selected candidates are shown with different colors. The annotator can choose any suggested option and the certainty degree of this election, or accept the solution given by the system.



Texto original

Hay veces que el cuerpo nos engaña por el puro placer de engañarnos . *0* Dicen , por_ejemplo , que algunas tardes a los mancos les duele el brazo que *0* no tienen : es curiosa esa nostalgia que el organismo expresa por el miembro perdido . En las novelas de aventuras el que engaña es el sentido de la vista , que hace que el explorador desorientado descubra un oasis donde sólo hay una duna y los restos de un camello muerto . Los espejismos , según el diccionario , no son más_que meras ilusiones ópticas , pero yo , que nunca he sido personaje de novela ni compañero de aventuras de Lawrence_de_Arabia , lo más parecido a un espejismo que *0* he visto son esos bichitos blancos con forma de paramecio que culebrean sobre el forro oscuro de los párpados . *0* Se llaman fosfenos , y es divertido seguirlos mientras *0* sobreandan en la inestable marea de los ojos . A_lo_mejor esos minúsculos peces abisales existen de_verdad dentro_de nosotros y lo que ocurre es que sólo en raras ocasiones *0* podemos verlos . Cuando *0* no los vemos sino_que *0* oímos en nuestro interior su sordo zumbido de

Opciones

los mancos
el cuerpo
algunas tardes
veces que el cuerpo
el puro placer de eng

Estado: ● standby ○ certain ○ uncertain

Palabra: les

Solucion: ???

Propuesta: los mancos

Anaforas □ **Ver solo no resueltos**

```
*0* ===>  Sin solucionar
les ===>  Sin solucionar
*0* ===>  Sin solucionar
*0* ===>  Sin solucionar
*0* ===>  Sin solucionar
Se ===>  Sin solucionar
```

Figure 2: 3LB-RAT anaphoric annotation tool.

6 Conclusions

The main contribution of this paper is the application of semantic information to a manual anaphora annotation process, based on the semantic relation between the anaphoric element and its antecedent at discourse level.

The semantic and anaphoric annotation scheme of the Spanish corpus Cast3LB has been presented, and how anaphoric annotation has been improved with the semantic information annotated in previous steps. The annotation process is based on the help of an anaphora resolution system: first, the system detects the anaphora and its antecedent, and then the human annotator checks the correctness of the automatic annotation process and solves possible mistakes. The system uses all the linguistic information previously annotated in the corpus, including the semantic information, in order to evaluate the semantic compatibility between the antecedent and the verb of the sentence in which the anaphora appears.

Acknowledgements

The authors would like to thank Belén and Raquel for their work in the manual annotation process.

References

M. Civit, M^a. A. Martí, B. Navarro, N. Bufí, B. Fernández, and R. Marcos. 2003. Issues in the Syntactic Annotation of Cast3LB. In *4th International Workshop on Linguistically Interpreted Corpora (LINC03), EACL03*, Budapest.

M. Civit. 2003. *Criterios de etiquetación y desambiguación morfosintáctica de corpus en Español*. Sociedad Espaola para el Procesamiento del Lenguaje Natural, Alicante.

R. Gaizauskas and K. Humphreys. 1996. Quantitative evaluation of coreference algorithms in an information extraction system. In S. P. Botley and A. M. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*, pages 143–167. John Benjamins, Amsterdam.

L. Hirschman. 1997. MUC-7 coreference task definition Message Understanding Conference Proceedings.

A. Kilgarriff and M. Palmer. 2000. *Computer and the Humanities. Special Issue on SENSEVAL*, volume 34.

A. Kilgarriff. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language. Special Use on Evaluation*, 12(4):453–472.

J. Lyons. 1977. *Semantics*. Cambridge University Press, London.

G. A. Miller. 1990. Wordnet: An on-line lexical database. *Intenational Journal of Lexicography*, 3(4):235–312.

R. Mitkov, R. Evans, C. Orasan, C. Barbu, L. Jones, and V.Sotirova. 2002. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, Lancaster, UK.

R. Mitkov. 2002. *Anaphora resolution*. Pearson, London.

B. Navarro, M. Civit, M^a. A. Martí, B. Fernández, and R. Marcos. 2003. Syntactic, semantic and pragmatic annotation in Cast3LB. In *Proceedings of the Shallow Processing of Large Corpora. A Corpus Linguistics WorkShop*, Lancaster, UK.

H. T. Ng and H. B. Lee. 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California.

M. Saiz-Noeda. 2002. *Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español*. Ph.D. thesis, Universidad de Alicante, Alicante.

N. Sebastián, M^a. A. Martí, M. F. Carreiras, and F. Cuetos. 2000. *2000 LEXESP: Léxico Informatizado del Español*. Edicions de la Universitat de Barcelona, Barcelona.

P. Vossen. 1998. *A Multilingual Database with Lexical Networks*. Kluwer Academic Publisher.