# Supporting anaphor resolution in dialogues with a corpus-based probabilistic model

**Marco Rocha**
School of Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QH, U.K.
marco@cogs.susx.ac.uk

## Abstract

This paper describes a corpus-based investigation of anaphora in dialogues, using data from English and Portuguese face-to-face conversations. The approach relies on the manual annotation of a significant number of anaphora cases - around three thousand for each language - in order to create a database of real-life usage which ultimately aims at supporting anaphora interpreters in NLP systems. Each case of anaphora was annotated according to four properties described in the paper. The code used for the annotation is also described. Once the required number of cases had been analysed, a probabilistic model was built by linking categories in each property to form a probability tree. The results are summed up in an antecedent-likelihood theory, which elaborates on the probabilities and observed regularities of the immediate context to support anaphor resolution by selecting the most likely antecedent. The theory will be tested on a previously annotated dialogue and then fine-tuned for best performance. Automatic annotation is briefly discussed. Possible applications comprise machine translation, computer-aided language learning, and dialogue systems in general.

## 1 Introduction

The emergence of corpus-based approaches brought to the fore the importance of extensive records of real-life language. The technique of corpus annotation and the use of statistical measures are standard research tools in corpus-based approaches. This paper presents a study which relies on corpus annotation to describe anaphoric phenomena in two languages - English and Portuguese. The investigation concentrates on dialogues. The London-Lund Corpus is the source of English data, whereas the Portuguese data come from a corpus collected especially for the purposes of this research.

Fligelstone's (Fli92) study on anaphora bears important similarities to the present one, as it also uses an annotation to describe features of anaphoric phenomena. The annotation created for the present study draws on some of the ideas which guide Fligelstone's, but it is quite distinct in both form and content. Biber's (Bib92) systematic use of statistical techniques to explore corpus data, together with the broad concept of referring expressions adopted, was also influential in shaping choices made for this project.

Having in mind Biber's non-restrictive approach, anaphora is defined, for the purposes of this research, as the relationship between a term - called the **anaphor** - which must be linked to an explicit or inferable element in the discourse - called the **antecedent** - in order to successfully accomplish semantic interpretation. All types of anaphors are annotated, including pronouns, noun phrases, verb phrases, and all elliptical phenomena.

A number of studies on anaphora attempt to incorporate the notion of topic, focus, or centre to the analysis of anaphora (see, among others, (Sid86), (Fox87)), leading to the discussion of ways to track topic - under any of the various names - in discourse (among many others, (Rei85), (GS86) and (GJW95)) and to relate topicality to anaphor resolution. The research described here is no exception. In order to assess the importance of topicality for anaphor resolution, it was decided that topic structure should be made an integral part of the investigation, and, consequently, encoded in the annotation.

The notion of topic is, however, notoriously difficult to deal with (see (BY83) for an extensive discussion). A routine dialogue contains a number of discourse entities, typically expressed by noun phrases, which, to mention a few possibilities: may retain a salient status throughout the whole dialogue; may pop in and fade out any number of times; may pop in once and fade out for good; may pop in and subdivide into subordinate topics, then fade out and then return; and several other possible combinations and interactions. Moreover, real-life conversations often cannot be summed up in terms of a title-like global topic in any easy way.

The study thus aimed at a working definition for the different levels of saliency so as to make the notion of topicality useful for the purpose of anaphor resolution. A set of categories was created to classify discourse entities into **topical roles** which cover the various levels of saliency. Global and local topics for a given dialogue had to be established a priori, independently of the analysis of anaphoric relations, so as to avoid circularity, as pointed out in (Fox87), although subsequent adjustments may consider discourse information related to those anaphoric relations.

Procedures to identify each one of the topical roles were spelled out as precisely as possible, having in mind that a measure of flexibility was necessary. The picture of topicality thus obtained does not claim to be any more than part of the truth. However, the assignment of topical roles to discourse entities is claimed to be an effective way of supporting anaphor resolution by keeping track of salient discourse entities.

## 2 The annotation

The annotation is manually entered by the analyst in separate lines inserted in a machine-readable transcript of a dialogue. Lines with one asterisk at the beginning contain information about the topicality structure. A one-asterisk line is inserted at the top of the transcript, defining which discourse entity is to be considered the global topic - called the **discourse topic** and represented by the code **dt** in the annotation - for the dialogue. The procedure to select the discourse topic draws on the work in (Hoe91) and involves a number of steps based on frequency, distribution, position of first token, and semantic adequacy for the role. In case there is a radical and stable change of topic within the dialogue, the dialogue is split into two fragments, each one with its own discourse topic.

Each local topic - called a **segment topic** and coded as **st** - is specified in one-asterisk lines inserted at the beginning of the segment in question. New segments introduce new local topics. The procedure to identify a new topic is based on the work on discourse analysis described in (Sin93) and in (Sin92), making use of concepts such as prospection and encapsulation. Each new utterance is assessed on the basis of these coherence devices to determine whether it introduces a new topic or not.

It is necessary further to divide the dialogue into subsegments with distinct subtopics, called **subsegment topics** and coded **sst**. These are subordinate local topics within a segment. Subsegment topics are also specified in one-asterisk lines by means of an ss mark that distinguishes them from segment topics (marked s). Therefore, the procedure used for segmentation must not only identify a new topic but also distinguish a local topic from a subordinate local topic.

Segments and subsegments are sequentially numbered as they appear in the dialogue. In case a previously developed segment or subsegment topic becomes the current topic again, the code **r** is placed before the segment or subsegment code to signal it is a resumptive segment or subsegment. Subsegment codes are followed by a slash which is in turn followed by the code for the segment which contains the subsegment (see example (1) below).

The procedures used to assign topical roles to discourse entities aim to be as objective as possible, so that different analysts should come to the same conclusions concerning topical roles in a given dialogue. The procedures constrain choices, but the analyst must use a measure of discretion to make final decisions. A full description of the procedures, as well as the complete listing of codes used in the annotation scheme, can be found in (Rocng).

Once the topicality structure of the dialogue has been fully defined, each case of anaphora is annotated according to four properties. The first property is the **type of anaphor**. The categories used combine: word classes, such as **subject pronoun** (coded as **SP** in example (1) below); phrase structure concepts, such as **noun phrase**, marked **FNP** in (1); and anaphora-world definitions, like **one-anaphora**, which appears as **One_an** below. The code is entered in a line with two asterisks at the beginning, inserted under the anaphor classified.

Notions like zero anaphor or zero pronoun are not included in the set of categories employed to classify types of anaphor. The word which triggers a search for an antecedent is annotated as an anaphor. A verb which appears without one or more of its essential complements requires the identification of these complements from ongoing discourse and is consequently annotated as an anaphoric verb. This decision is particularly important for the annotation of the Portuguese data. The twenty-seven categories used in the analysis of the English sample were grouped into three umbrella categories. Frequencies for these umbrella categories are shown in Table 1 below:

Table 1: Frequencies for types of anaphor

| | Frequency | Percent |
|---|---|---|
| Pronouns | 1579 | 51.1 |
| Verbs and adverbials | 318 | 10.3 |
| Nominals | 1193 | 38.6 |
| Total | 3090 | 100.0 |

The three remaining properties are entered in a line with three asterisks at the beginning inserted under the two-asterisk line with the code for the type of anaphor. A semicolon separates the code for each property. An example of annotated text is shown below:

```
(1)
B: well I think probably what Captain Kay
**                            FNP
***            ex_222; dthel; LR;
must have said was a will is legal if it's
**                                    SP
***                  ex_224; dthel; FtC;
 witnessed on the back of an envelope

*  ss4/s38  'Captain's personal witnessing'

A: w- did he say that
**        SP
***       ex_222; thel; FtOp;
    he had personally witnessed one
**  SP                          One_an
*** ex_222; thel; FtCCh;  ex_1; dt; SetMb;
B: well I could have been
   I could have been wrong there
**                         AdvP
***          ex_116; p_sst; CK;
```

The first property to have the corresponding code inserted in the three-asterisk line is the **type of antecedent**. The antecedent for the anaphor in question is classified according to the explicit/implicit dichotomy, using the marks **ex** and **im** followed by a number which identifies the referent in a list. However, it is a policy of the study to annotate every token of third-person personal pronoun, as well as all demonstrative pronouns, regardless of the fact that they may be nonreferential, and thus not a case of anaphora strictu sensu. A third category was created for the cases of nonreferential pronouns - typically *it* or *that*. Frequencies for the English sample are shown in Table 2 below:

Table 2: Frequencies for types of antecedent

| | Frequency | Percent |
|---|---|---|
| Explicit | 2562 | 82.9 |
| Implicit | 412 | 13.3 |
| Nonreferential | 116 | 3.8 |
| Total | 3090 | 100.0 |

The second slot in the three-asterisk line contains code for the property called the **topicality status of the antecedent**, which uses the topical roles defined for topic tracking to classify the antecedent of the anaphora case in question. An antecedent which is not one of the topics is a discourse entity associated to one of the topics. If it is associated locally to the segment topic, it is classified as a **thematic element**. A thematic element may have a cross-segment saliency, in which case it is called a **discourse thematic element**. The latter typically include the participants in the dialogue, other important agents and also objects associated to the discourse topic.

Antecedents can also be discourse chunks. They are classified as predicates of the entity with a topical role to which they are most strongly related. The various categories used to assign a topicality status to antecedents were grouped as global (discourse) roles, local (segment) roles, or sublocal (subsegment) roles. A fourth category - namely, **focusing device** - is used to classify the cases of anaphors with no antecedent (nonreferentials) or with antecedents which were too vaguely implicit for an accurate assessment in terms of topical role. Frequencies for the English sample are shown in Table 3 below:

Table 3: Frequencies for topical roles

| | Frequency | Percent |
|---|---|---|
| Local topical roles | 1298 | 42.0 |
| Global topical roles | 1068 | 34.6 |
| Sublocal topical roles | 585 | 18.9 |
| Focusing devices | 139 | 4.5 |
| Total | 3090 | 100.0 |

The fourth property is an attempt to encode psycholinguistic information for anaphor resolution. The observation of corpus data revealed that the classification into types of anaphor - first property - did not cover important processing information. Different strategies are needed to resolve the same type of anaphor - and often the same anaphoric word or phrase - in different contexts. Syntactic information - as codified in an algorithm like the "naive" algorithm in Hobbs' (Hob86) - may suffice to resolve a given occurrence of *it*. However, another token of the same word may demand rather complex discourse processing, bypassing a number of candidates to reach the correct antecedent. A large number of categories were used to classify tokens according to **processing strategy**. They were grouped as shown in Table 4 below with the respective frequencies for the English sample.

Table 4: Frequencies for processing strategies

| | Frequency | Percent |
|---|---|---|
| Lexical processes | 1095 | 35.4 |
| Discourse processes | 503 | 16.3 |
| Collocations | 279 | 9.0 |
| Syntactic processes | 1213 | 39.3 |
| Total | 3090 | 100.0 |

## 3 The probabilistic model

The frequency counts yielded by the annotation work - shown in the previous section - were used to build a probabilistic tree which is a model of the anaphora world as described by the annotation scheme. The root of the tree is a category in the variable named **type of anaphor**. The choice bears in mind the possibility of automatic annotation. Given a POS-tagged dialogue, it should not be difficult to map the tags into the categories used to classify the type of anaphor.

It was necessary then to decide which variable should occupy the next level in the tree. In order to make an informed choice, cross-tabulations for each possible combination of two variables were produced, together with a chi-square test and two non-chi-square-based association measures. Significance was achieved in all cases, but association was not very strong, except for the relation between type of anaphor and processing strategy (Goodman and Kruskal tau = 0.41). The Goodman and Kruskal tau is an association measure based on the notion of proportional reduction of error. The value thus means that, once the distribution for type of anaphor is known, the chances of predicting the processing strategy correctly are forty-one percent higher.

Other factors pointed to the processing strategy variable as the best candidate for the second level of the probability tree. The other two variables classify the antecedent. Thus, it is impossible to be sure of the correct category classification before actually identifying the antecedent. This means that, although the type of antecedent can occasionaly be predicted on the basis of the anaphor type, it will not be possible to offer more than a probability for each category in most cases. On the other hand, the processing strategy can be safely predicted on the basis of the anaphor type in at least one case, namely, if the processing strategy relies on knowledge of collocations. These collocations contain words such as *it* or *that* which function in a distinct way when appearing in phrases such as *that's it* or *I mean it*. Collocations can be identified by simply checking a list which has been prepared as the annotation work progressed.

The nodes on the second level of the tree are the categories which classify the processing strategy. Each branch of the tree is assigned two values. The first one is the probability for that particular branch within the universe of the node immediately above, while the second one is the probability for the whole branch all the way to the root, that is, in relation to the total sample. Thus, given that the anaphor is a pronoun, the probability that it will be resolved by means of **lexical processing** - meaning knowledge associated with the semantics of the anaphor - is 0.01267, which is rather small. In relation to any anaphor, the probability that it will be a pronoun resolved by means of lexical processing is 0.00647, which is extremely small. However, it is different from zero and must be taken into consideration.

The subsequent level in the tree can be occupied by any of the two remaining variables. However, it was decided that probabilities should be calculated for all possible combinations of categories across the variables. Once the frequency counts had been obtained, a program was written which calculates probabilities for every combination in relation to the immediately higher node and for the total in all possible orderings of the variables. In spite of the fact that placing the processing strategy before the other two is clearly more economic, there may be one type of anaphor for which this is not true. All options are thus available for use in building the antecedent-likelihood theory.

The probabilistic model is the mainstay of the theory, but the collocation list and other regularities observed also play an essential role. For instance, the few cases classified as pronouns resolved by lexical processing were looked into in search of a feature that could be the clue for pronoun resolutions based on lexical processing. Probabilities for the ungrouped categories were also calculated and are a source of useful information as well. The next section describes how these various inputs are combined to support anaphora resolution.

## 4 Building the theory

Once the probabilities for every combination of categories across the variables had been worked out, the task then was to put these numbers to good use. In the case where pronouns are the root of the probability tree, the results for processing strategy are as shown in Table 5 below.

Table 5: Processing strategies for pronouns

| | Frequency | Probability |
|---|---|---|
| Lexical processes | 20 | 0.012 |
| Discourse processes | 398 | 0.252 |
| Collocations | 217 | 0.137 |
| Syntactic processes | 944 | 0.597 |
| Total | 1579 | 1.000 |

If these results are compared to the percentages in Table 4, some important differences emerge. There is a steep decline in the number of anaphors resolved by means of lexical processes. This is not surprising. Lexical processes are an umbrella category grouping strategies such as **world knowledge** and **lexical repetition**. These strategies are typical of resolutions related to anaphoric nonpronominal noun phrases, as they rely on the semantic content of the anaphor itself to identify the correct antecedent. As pronouns characteristically have low semantic value, it is in fact surprising that any of them are resolved by such means at all.

All other three categories show increases in relation to the percentages in Table 4, but syntactic processes present the highest increase. One of the strategies grouped under syntactic processes is the **first-candidate** strategy, which may be described as an adaptation of Hobbs' "naive" algorithm (see (Hob86) to spoken language, since it searches for the first appropriate noun phrase in the ongoing discourse and selects it as the antecedent on the basis of agreement and syntactic constraints.

The most frequent processing strategy within syntactic processes is the **first-candidate chain**. This confirms Biber's (Bib92) findings about the importance of chains in conversations, but it tones down optimistic expectations of easy anaphor resolution. Chains do not necessarily start with an anaphor resolved by a first-candidate strategy, although many of them do. Consequently, the actual identification of the antecedent may still need to employ one of the less straightforward strategies. The two first-candidate strategies together account for almost all cases of syntactic processes in pronouns.

The list of collocations collected during the annotation process shows that, within the pronoun category, the personal pronoun *it* and the demonstratives *this* and *that* are the only tokens which appear in collocations. There is no need to check the collocation list when the pronoun being resolved is not one of the above. Virtually all collocations entail a resolution for the anaphors they contain. Once identified, the collocation can therefore be associated to a distinct way of handling the anaphor.

Discourse processes are strategies that demand more complex information which cannot be obtained by checking a collocation list or analysing the semantic content of the anaphor. A first-candidate search will also fail in these cases, as the correct antecedent is not the first candidate available, either straightforwardly or in a chain. The typical case is the pronoun reference which bypasses the first candidate in spite of the fact that it is an appropriate one, if only agreement and syntax are considered. An example is given below:

```
(2)
B: I mean what difference could it make
   to the directors of Unilever that
   their shares had got down from say
   eighty to fifty or whatever it is
A: well in the present circumstances
   not very much because I mean
   everything has gone down but of course
   if they are consistently low
   it makes them more difficult
   it makes it more difficult for them
   to raise money
```

The second occurrence of *them* - the first one is part of a false start - is to a certain extent ambiguous, as the antecedent might be said to be either *directors of Unilever* or *Unilever*, although understanding is not much affected by choosing one or the other. What is important is that the antecedent is not *shares* and thus there is no chain of reference. The first candidate *they* has to be bypassed, as well as *present circumstances*, in order to identify the correct antecedent.

The phrase *to raise money* has to be semantically processed before the anaphor can be successfully resolved. Information yielded by the syntactic structure, lexical content of the anaphors, or knowledge of collocations will not achieve the correct identification of the antecedent. As the resolution involves knowledge only available after processing discourse in full, this strategy is named **discourse knowledge**. The use of lexical clues from the immediate context and the topical roles of candidates are of crucial importance for the correct resolution of this kind of anaphor.

Other strategies grouped under discourse processes include: **secondary reference**, which is the use of first and second person pronouns in speech reported verbatim to refer to persons previously mentioned in the dialogue; **distant anaphora**, which are pronouns with very distant antecedents - over fifty tone units - but without competing candidates; pronouns which conjoin referents in a set, called **set creation**; reference to an element within a set, called **set member**; and the cases of antecedentless anaphors (see (Cor96)), in which the processing strategy is called **deixis**. The categories grouped as discourse processes may be seen as the particularly complex strategies for anaphor resolution.

The example above also contains four tokens of *it*. Three of them can be resolved by using a more sophisticated version of collocational knowledge. The first one is in a *make no difference* collocation. The observation of corpus data shows that the *it* in such collocations has an explicit clausal or sentential antecedent in all cases found. It also reveals that the reference is cataphoric whenever "make" is the main verb in a sentence with a subordinate *that*-clause. Furthermore, this *that*-clause is the antecedent in all occurrences of the kind.

The collocation list has thus an entry such as:

**it X-make difference to Obj that-clause**

- cataphoric it (Subj)
- antecedent = that-clause

This sort of knowledge is extended to cleft sentences, adding to the collocation list an entry like:

**it X-be SubjC that-clause**

- cataphoric it (Subj)
- antecedent = that-clause

In order to resolve the second and third tokens of *it*, the entry to be accessed in the collocation list is:

**it X-VERB Obj1 Adj for Obj2 NF-clause**

- cataphoric it (Subj)
- antecedent = NF-clause
- if VERB = make and Obj1 = it
  - it (Obj1) nonreferential

The *X*- symbol means any inflected form of the verb, optionally including tense, aspect and modality. The major structures of the language, such as affirmative, interrogative and negative forms, are also assumed as included in the entry. The other symbols in the entries above stand for subject (Subj), subject complement (SubjC) object (Obj), adjective (Adj) and nonfinite (NF).

The entries in the collocation list are related to specific pronouns. As mentioned before, *it* is the only personal pronoun to appear in collocations with a pattern of regular resolution. It is reasonable to think, thus, that other patterns may emerge if the categories in the anotation scheme are individually analysed out of the umbrella categories. Although the grouping was very useful for the significance and association tests, the antecedent-likelihood (AL) theory requires a return to the original categories, as well as the analysis of individual pronouns.

Suppose then that a dialogue tagged using the tagset in (Sam95) is being analysed according to the AL theory in order to resolve anaphors. A word tagged as PPH1 is a token of *it*. Suppose furthermore that this token of *it* has been identified as an object pronoun by means which need not be discussed here. The header for the word in the AL theory is:

- syntactic process = 0.729
- discourse process = 0.151
- collocation = 0.080
- lexical process = 0.013

If these numbers are compared to the numbers for pronouns as a whole, there is a substantial increase in the number of anaphors resolved by syntactic processes. The probabilities for resolutions which rely on knowledge of collocations and on discourse processes decrease, whereas the probability for lexical processes remains equally low. The reduction in collocation-related strategies can be explained. The number of collocations in which *it* is an object pronoun is much smaller. Moreover, cleft sentences are the most common collocation, and *it* is a subject pronoun in these sentences. The decrease in resolutions by means of discourse processes is caused by the fact that demonstratives have been taken out.

The next step is to match the tone unit in which the token occurs with the entries in the collocation list. If there is a match, the path to resolution is spelled out in the entry. If there isn't, the next step is to eliminate rare processing strategies which are only needed in special cases. One way to do that is to use the strategy with the highest probability to select a tentative antecedent and check the antecedent against information in the theory. If no appropriate referents are found, not even one which fits poorly, it must be one of the special situations. In the case of *it*, the two first-candidate strategies are by far the most probable and rarely fail to produce an antecedent. Understandably, all cases in the sample in which both did fail are tokens at the very beginning of the dialogues in question.

The only possibility then is that the anaphor is one of the rare cases of resolution by means of lexical processes. Shared knowledge allows the participants to identify an antecedent that has not been mentioned because in the situation where the conversation occurs, *it* can only mean one thing. It is a rare but interesting case for dialogue systems in which the same user is expected to have more than one session. The history of communications between man and machine would have to be available in order to allow resolution, as it is the anaphor that introduces the discourse entity in the dialogue.

In all cases in the sample, participants only introduce discourse entities in this way when they are central to the conversation yet to take place and thus have highly salient global topical roles. The antecedent is obviously implicit. The AL theory for *it* as an object pronoun specifies then:

**check collocation list**

- if no match found

**select first appropriate candidate**

- if no appropriate candidate found
- beginning of dialogue ?
- if not no record
- if yes lexical process; shared knowledge
- discourse topic or discourse thematic element in all cases
- implicit in all cases

Resolutions which require discourse processes are the most difficult to identify, particularly those cases in which the first candidate is not the correct antecedent and must be bypassed for a different one, as in example (2) above. However, antecedents requiring this sort of processing strategy for identification are usually highly salient elements. Moreover, a lexical clue of some kind is often present in the context.

In the case of both object and subject pronouns, the verb to which they are attached is of great importance. The provisional antecedent may be ruled out by selectional restrictions. It seems also important to have a record of verbs associated to discourse entities, as they are likely to be referred to as arguments of the same verb or of a similar one. Related adjectives and noun phrases attached to the same verb should also be examined. If the provisional antecedent has never appeared as an argument of the verb to which the anaphor is attached, the possibility of bypassing it should be considered. If bypassing it selects a highly salient entity, such as the discourse topic or a high-frequency discourse thematic element, and this entity has appeared as an argument of the verb in question, the resolution by disocurse knowledge is probably the best choice. Thus, the AL theory for *it* as an object pronoun proceeds as below:

- if an appropriate candidate found

**check selectional restrictions of verb**

**check history of verb in dialogue**

**check associated adjectives and noun phrases**

- if the antecedent fits, accept it
- if the antecedent doesn't fit

**select next candidate**

**repeat checks**

- if the antecedent fits

**check topical role**

- if dt, dthel or st

**bypass previous candidate**

The AL theory is still being finalised. When completed, it will contain systematised records like those above for all types of anaphor. It will be then tested on a previously annotated dialogue which has not been included in the training sample. Results will be evaluated according to two standards: the percentage of correct antecedents identified by the single or highest-probability choice selected by the theory when applied to a case; and the percentage of correct antecedents identified when lower-probability choices are also considered. The test will assess the efficacy of the theory and will also expose overlooked shortcomings.

## 5 Future developments

This paper presents results for the English sample only. The same set of categories is used for the annotation of dialogues in Portuguese. However, some types of anaphor only have tokens for one of the languages. For instance, the type of anaphor *one-anaphora* does not occur in Portuguese. One of the interesting developments to be explored, once the analysis of both samples is completed, is the contrastive analysis of results. A database of aligned discourse environments related to anaphoric phenomena - covering linguistic information at all levels - could be produced, providing guidance for applications such as machine translation and computer-aided language learning. If automatic annotation can be at least partially accomplished, the scheme may prove its worth in practical applications, including those which involve only one of the two languages, such as dialogue systems.

Automatic annotation using this scheme is a daunting task, particularly because of the need to identify the discourse entities selected for the topical roles, as procedures ultimately require a decision by the analyst. Other problems not discussed in this paper, such as the identification of discourse-chunk antecedents for the resolution of demonstrative pronouns, are also very difficult. Nonetheless, the approach seems worth pursuing precisely because the hardest cases are not left out. The inclusion of variables for topical roles and processing strategy represents an attempt to deal with difficulties which have been often avoided in studies on anaphora.

## 6 Acknowledgment

The project is fully funded by the Conselho Nacional de Desenvolvimento Cientifico e Tecnológico under grant no. 200608-92/4.

## References

Douglas Biber. Using computer-based text corpora to analyse the referential strategies of spoken and written texts. In Jan Svartvik, editor, *Directions in corpus linguistics*, pages 215–252, Berlin and New York, 4-8 August 1991 1992. Nobel Symposium 82, Mouton de Gruyter.

Gillian Brown and George Yule. *Discourse analysis.* Cambridge University Press, Cambridge, 1983.

Francis Cornish. Antecedentless anaphors: deixis, anaphora, or what? Some evidence from English and French. *Journal of Linguistics*, 32:19–41, 1996.

Steve Fligelstone. Developing a scheme for annotating text to show anaphoric relations. In *New directions in English language corpora: methodology, results, software development*, number 9 in Topics in English Linguistics, pages 153–170. Mouton de Gruyter, Berlin and New York, 1992.

Barbara Fox. *Discourse structure and anaphora.* Cambridge University Press, Cambridge, 1987.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

Barbara Grosz and Candace Sidner. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July-September 1986.

Jerry Hobbs. Resolving pronoun references. In B.L. Webber, Barbara Grosz, and K. Jones, editors, *Readings in Natural Language Processing.* Morgan Kaufmann, Palo Alto, CA., 1986.

Michael Hoey. *Patterns of lexis in text.* Oxford University Press, Oxford, 1991.

Rachel Reichman. *Getting computers to talk like you and me.* MIT Press, Cambridge, MA, 1985.

Marco Rocha. A description of an annotation scheme to analyse anaphora in dialogues. Technical Report 427, University of Sussex - School of Cognitive and Computing Sciences, Brighton, 1997 (forthcoming).

Geoffrey Sampson. *English for the computer.* Clarendon Press, Oxford, 1995.

Candace Sidner. Focusing in the comprehension of definite anaphora. In Karen Jones Barbara Grosz and Bonnie Webber, editors, *Readings in natural language processing.* Morgan Kaufman, Palo Alto, CA, 1986.

John Sinclair. Priorities in discourse analysis. In R. Coulthard, editor, *Advances in Spoken Discourse Analysis.* Routledge, London, 1992.

John Sinclair. Written discourse structure. In J. Sinclair, M. Hoey, and G. Fox, editors, *Techniques of description: spoken and written discourse: a festschrift for Malcolm Coulthard.* Routledge, London, 1993.