A description of an annotation scheme to analyse anaphora in dialogues

Marco Antonio Esteves da Rocha

427

February, 1998

ISSN 1350-3162

UNIVERSITY OF
SUSSEX
AT BRIGHTON

Cognitive Science
Research Papers

Acknowledgements

I am indebted to Geoffrey Sampson, who was a source of guidance and inspiration thorughout the whole investigation. I convey my appreciation also to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the full sponsorship of a project which included the designing of the annotation scheme, through the grant 200608-92-4. Several members of the faculty at the School of Cognitive and Computing Sciences provided useful comments and advice by a variety of means, ranging from questions and criticism at talks I gave in internal events to informal chats in the corridors. I am especially indebted to Nicola Woods and Gerald Gazdar for sharp and concise remarks, which were highly influential and nevertheless consumed a few minutes only. I am also grateful to Dr. Renato Veras, M.D., director of the UnATI, in Rio de Janeiro, for having allowed me to roam freely the premises of the university and the hospital, placing tape recorders in the hands of incredibly helpful staff at will. The high quality of the data collected for the Portuguese sample would have been unattainable without his cooperation. Errors and misconceptions are my exclusive responsibility.

For ne'er
Was flattery lost on a poet's ear:
A simple race! They waste their toil
For the vain tribute of a smile.
(Sir Walter Scott - The Lay of the Last Minstrel)

This is it !
(David A. Johnston (volcanologist) -
his last words when on duty on the 18th of May, 1980
at the time Mount St. Helen's volcano blew its top)

Because half a dozen grasshoppers under a fern make the field ring with their importunate chink, whilst thousands of great cattle, reposed beneath the shadow of the British oak, chew the cud and are silent, pray do not imagine that those who make the noise are the only inhabitants of the field; that, of course, they are many in number; or that, after all, they are other than the little, shrivelled meagre, hopping, though loud and troublesome, insects of the hour.
(Edmund Burke)

Foste ?
Fui.
Compraste ?
Comprei.
Me diz quanto foi.
Foi quinhento-réis.
(Children's wordplay)

Eu vou dar a despedida
Como deu o bacurau
Uma perna no caminho
Outra no galho de pau
(Folk rhyme)

"É bom pensar, sonhar consola."
"Consola, talvez; mas faz-nos também
diferentes dos outros, cava abismos
entre os homens..."
(Lima Barreto - O Triste Fim de Policarpo Quaresma)

A description of an annotation scheme to analyse anaphora in dialogues

Marco Antonio Esteves da Rocha

Abstract

This paper describes an annotation scheme designed for the analysis of anaphoric relations in dialogues. The scheme was developed by annotating a relatively large number of anaphora cases in English and Portuguese, using dialogue corpora. The corpora used as sources of data were the reformatted version of the London Lund Corpus, as stored in the School of Cognitive and Computing Sciences, at the University of Sussex, for the dialogues in English; and a corpus of dialogues in Portuguese collected for the purposes of this research, named the Rio de Janeiro Clinical Dialogues Corpus. The annotation scheme is intended as an analytical tool which attempts to show the relations between anaphors, as they appear verbatim in spoken language, and the required processing for the identification of the antecedent. Each case of anaphora is classified according to four properties, namely: type of anaphor, type of antecedent, topical role of the antecedent, and processing strategy. The set of categories used to classify the anaphora cases according to these properties is described in the paper. The rationale underlying the choice of properties is also discussed. The annotation scheme is thought to be useful for the purposes of encoding discourse relations in text, as well as a way of supporting anaphora resolution in natural language processing.

University of Sussex
School of Cognitive and Computing Sciences
Falmer, Brighton - BN1 9QH
marco@cogs.susx.ac.uk

Contents

Chapter 1

Introduction

The quotations at the beginning of the paper are meant to illustrate the pervasiveness and the wide range of forms under which the linguistic phenomenon studied appears in everyday language. Thus, the first excerpt from Sir Walter Scott could be seen as a relatively formal or at least carefully planned passage of written language, as it is, in fact, poetry. It is rather short, but it contains two anaphors, one of which is a nonpronominal noun phrase, and the other is a pronoun, the word class most overtly associated with the idea of anaphora. Nonetheless, the interpretation of these two anaphors is not so simple a matter as the unadorned beauty of the verses might suggest.

The antecedent for the noun phrase *a simple race* is *a poet*, which is introduced as a genitive noun in the noun phrase that immediately precedes the anaphoric one. It is therefore necessary for the interpreter to know that the noun *race* usually refers to humans and not to parts of the body or to acts such as *flattery*. This is followed by a plural third-person pronoun, which assumes that the interpreter understands the collective meaning of *race* and is thus perfectly capable of dealing with the apparent lack of number agreement. All this must be accomplished while the altered sense of *race* and other adjustments are made, so as to actually understand the passage.

The vulcanologist's words are even more mysterious. One might play with themes such as the immense power of natural forces or death, but it would not be easy to assign precise referents to the pronouns, although we can all understand the epiphany-like feeling they convey. A number of skills, ranging from complex lexical semantics to effects of syntactic parallelism, are required in order to distinguish *the insects of the hour* from *the great cattle* in the quotation extracted from Burke's pages. Words like *ring*, *chink* and *noise* must be semantically interpreted as signals of coreference, and the sequence of third-person plural pronouns forms a chain which the interpreter is expected to construct and keep attached to the correct antecedent.

As to the Portuguese quotations, they add some extra fascination to the object of research. The last one is taken from a Brazilian early twentieth-century writer who is noted for a polished style which nevertheless manages to capture the speech of middle-class Rio de Janeiro inhabitants of the time. Coração dos Outros (CO below) is talking to the main character, Policarpo Quaresma (PQ). The glosses below help the contrast with English. See Appendix A for a list of the conventions used in the glosses.

(1) **CO:** É bom pensar, sonhar consola
gl: Is good think-INF, dream-INF consoles
tr: It is good to think, dreaming comforts

PQ: Consola, talvez; mas faz-nos também diferentes dos outros
gl: Consoles, perhaps; but makes-OBjP1stp also different from others

tr: It comforts, perhaps; but it also makes us diffferent from others

PQ: cava abismos entre os homens...
gl: digs abysses between the-MASCp men
tr: it digs chasms between men

Two aspects of the cross-linguistic contrast are easily discerned. The first one is that subjects can be omitted in Portuguese as a matter of course. Omitted subjects are identified by retrieving them from preceding discourse and, in some cases, by more complex means. The second aspect is that Portuguese does not have a neuter personal pronoun to match *it* in English, although neuter forms survive as demonstratives *isso* and *aquilo*. Sentences like the first one in the example above are thought of as an inversion which is frequent in copular constructions. References to inanimate objects can be made by using the personal pronouns *ele(s)* and *ela(s)*, but subjects seem to be more often omitted when they are not persons. The singer (S) in the folk rhyme adds a new element to the contrast.

(2) **S:** Eu vou dar a despedida
gl: I go give-INF the-FEMs goodbye
tr: I'm going to say goodbye

S: como deu o bacurau
gl: as gave-3rds the-MASC bacurau (a bird)
tr: as the bacurau did

S: uma perna no caminho
gl: one-FEM leg on-the-CONTR road
tr: one leg on the road

S: a outra no galho de pau
gl: the-FEM other-FEM on-the-CONTR branch of wood
tr: the other on the tree branch

The verb *dar*, which appears in the first line of the rhyme, is a ditransitive, taking one direct object — what is given — and one indirect object — to whom it is given. The token above is not used in the conventional sense. The usage of this particular verb is very flexible in Portuguese. In this case, it combines with the noun *despedida* to replace the verb *despedir-se*, which would be more natural but would spoil the metrics. Thus, the indirect object would not be expected in this case, but the direct object, introduced in the first line, is also omitted in the second line. The subject is postposed for the sake of rhyming with the last line.

The omission of the direct object is typical of spoken language, and it would not be tolerated in formal writing. The interpretation of anaphoric references relies heavily on the information conveyed by the argument structure of verbs for identification of the antecedents. This feature of the language is exploited in the children's wordplay as a kind of imagination game. The children are arbitrarily called C1 and C2 in the example below.

(3) **C1:** foste ?
gl: went-2nds
tr: Did you go ?

C2: fui

gl: went-1sts
tr: I did

C1: compraste ?
gl: bought-2nds
tr: did you buy it ?

C2: comprei
gl: bought-1sts
tr: I did

C1: me diz quanto foi
gl: me tell how much was-3rds
tr: tell me how much it was

C2: foi quinhento-réis
gl: was-3rds five hundred réis (old Brazilian currency)
tr: it was five hundred réis

The objects of the verbs are unknown, although the subjects are identifiable because of morphology. One may freely construe places where one would go and buy something which cost *quinhento-réis*, as the language is perfectly understandable and quite normal. The same sort of game could of course be played in English with pronouns in the appropriate places, but obviously this makes quite a difference in a referring system. The processing does not rely on indexes which signal the need to retrieve an antecedent, but on a verb with omitted arguments. The use of inflected verb forms instead of the pronoun-operator system used in English is also a noteworthy contrast at the root of important developments in spoken Portuguese.

The study of anaphoric phenomena in real-life dialogues involves thus a variety of forms of reference, instantiated by pronouns, noun phrases and verbs, and, more importantly, a variety of processing demands to which solutions must be found with the means available, that is, those provided by the dialogue itself and the linguistic knowledge of participants. The successful interpretation of spoken discourse requires linking all those different parts of speech to keep track of referents which appear and depart or stay as the dialogue develops. This must be accomplished in real time, a task which language users regard as trivial, but which draws on all levels of linguistic knowledge. The phenomenon is so ubiquitous in natural language under so many forms that defining the actual matter of study can be quite difficult.

No wonder, thus, that the examination of the large body of literature concerned with anaphora shows that the term is used to refer to a variable range of linguistic phenomena, although the third-person personal pronoun remains as the prototypical and most often investigated form of anaphora. As noted by Bosch [Bos83], the word **anaphora** was in many ways a handy solution for the problems identified with the term **pronominalisation**, since the literal sense of the word **pronoun** leads to unwanted assumptions. Pronouns are not simply a substitute for a noun or noun phrase that could be used instead. It seems counter-intuitive to understand first-person pronouns as substitutes for the name of the person who is speaking. The substitution approach also runs into difficulties for third-person pronouns in nonreferential uses and references to discourse chunks, as well as in many other cases (see [QGSL85], section 2.44).

On the other hand, it is evidently true that pronouns have little semantic content and therefore need to be somehow linked to other elements for successful semantic interpretation. These other elements must be either present in the discourse or inferable from what has been said. The surrounding physical environment and the situation are also crucial sources of information for the

necessary interpretation, especially in deictic uses[1]. Seen from a processing viewpoint, pronouns are words which refer to another retrievable element in an instance of discourse by standing in a special kind of relationship to this object of reference. Anaphora is therefore a name for a relationship or process in which a term — called the **anaphor** — in an instance of discourse is linked to an identifiable element — called the **antecedent** — in order to successfully accomplish semantic interpretation[2].

A desirable consequence of the change in terminology is the possibility of including phenomena which are not necessarily related to pronoun reference without misnaming them. Under the name of anaphora, pronouns can be understood as a manifestation of a far more general process: the use of a variety of linguistic devices to achieve cohesion, as defined in Halliday and Hasan [HH76]. Although pronouns remain as the most common object of analysis in research dealing with anaphora, several works sought to discuss other forms of anaphoric reference, such as anaphoric nonpronominal noun phrases and verbal ellipses (see Chapter 2). This expansion of the concept was often associated with research concerned with discourse phenomena, carried out not only by linguists, but also by psychologists and researchers in the fast-growing field of natural language processing.

The approaches which remained within the limits of sentence grammar — notably generative grammar — developed studies on syntactically-controlled anaphora. A large number of anaphora cases was thus kept out of the investigation as occurrences of 'pragmatically-controlled' anaphora with no place in linguistic theory. A practice of creating examples, instead of extracting them from observable data, was developed along with these approaches. This practice was justified by the belief that true linguistic knowledge was to be found outside everyday language as used in context for communication. The degree of abstraction which linguistic theory progressively reached made its formulations hard to recognise when confronted with everyday language. Consequently, for both psycholinguistics and natural language processing (henceforward NLP), such approaches do not offer solutions for many of the problems at hand, as these two fields cannot discard phenomena of everyday language.

The methodology of corpus linguistics offers an alternative to those researchers who felt uncomfortable with the distancing of linguistic theory from everyday language. It also brought hope to the serious difficulties experienced by NLP systems when confronted with real-life language. The amount of corpus-based work has grown steadily in recent years, although it is still relatively small in terms of discourse-oriented research. Differently from structural approaches, grammars and theories in corpus-based research are developed out of a thorough survey of occurrences for a given phenomenon as observed in everyday language. Created examples are not the rule, but rather the exception. All cases of the phenomenon in question are included, and statistical notions such as frequency and probability play a significant role in the formulation of theory.

This paper concentrates on the description of the annotation scheme created for the analysis of anaphora in corpora of dialogues in English and Portuguese. The corpora used as sources of data were the Reformatted London Lund Corpus (RLLC), for the English data, and the Rio de Janeiro Corpus of Clinical Dialogues (CDC-RJ), which was collected for the purposes of this research. The use of corpus annotations as research tools is now quite well established as a technique to analyse data related to any linguistic phenomenon observed in samples of real-life language. As a natural result of the relatively few corpus-based studies on discourse relations, there were virtually no choices of pre-existing annotation schemes to analyse anaphora. The only similar scheme found is discussed in Chapter 3, together with reasons for the creation of a distinct scheme. They ultimately boil down to the specific requirements related to spoken language and cross-linguistic analysis. It seems reasonable to believe that the scheme could be used to analyse anaphoric relations in languages other than English and Portuguese, notably those most closely related to them.

[1] See Bosch [Bos83] for a discussion of the distinction between anaphora and deixis.

[2] The terms **anaphor** and **antecedent** are commonly used to analyse cataphoric reference as well. They are also applied to the analysis of deictic references in this study.

Relative frequency of occurrence for the different types of anaphor is seen as an important aspect of anaphoric relations. Thus, it was decided that every form of anaphoric relation would be included in the scheme, including personal pronouns, independent and determinative possessive pronouns, demonstrative pronouns, anaphoric nonpronominal noun phrases, anaphoric adverbs of place, one-anaphoras, nominal ellipses, verbal ellipses of all kinds — including short answers — reflexives, and reciprocals. Despite the risk of excessively widening the scope, it was thought that the frequency results were crucial, especially because of the cross-linguistic analysis, as some types of anaphor may appear as a different type of anaphor in the other language. Annotating every form of anaphora allows the quantification of such cross-linguistic phenomena.

The annotation scheme is built on the assumption that third-person pronouns and demonstrative pronouns are invariably annotated, even when they are not truly anaphoric, that is, in those cases when they do not have a referent, such as in the weather constructions in English. It is also assumed that deictic uses of these pronouns will be annotated, as well as pronouns contained in collocations, which are a special form of anaphoric relation. The reason for the assumptions is again to allow the frequency of these tokens to be established, as well as the contextual features in which they occur, if any are regularly present. It is also assumed that first-person and second-person pronouns are not annotated when they appeared in their typical function, referring to the participants in the dialogue. They were annotated when they appeared in speech reported verbatim, referring to third parties.

The annotation is concerned with computer applications which might be developed on the basis of the findings. Thus, computational efficiency in terms of search-and-retrieval operations is kept in mind as one significant factor, shaping choices for the annotation features. The research paper is organised as follows: the next chapter briefly reviews works which are related to choices made for the definition of the annotation scheme; the third chapter discusses the methodological aspects of a corpus-based analysis of anaphora and concludes with the rationale which underlies characteristics of the annotation, such as the inclusion of elements to analyse topicality and processing; the fourth chapter describes the annotation scheme in full detail, with as many examples as needed to make distinctions clear.

Chapter 2

Brief overview of the literature

This review is obviously not meant to be exhaustive, as this would be a Herculean task regarding research on anaphora. The problem is compounded by the need to discuss research aimed at discourse analysis, as well as discourse modelling in NLP systems, since they are intrinsically related to anaphoric phenomena. Theoretical works which rely on a structural sentence-grammar approach to spell out constraints on anaphoric reference are not mentioned here (but see Reinhart [Rei83] for a complete discussion of such an approach). In fact, the review discusses a small number of studies which play a salient role in the definition of the conceptual framework underlying the choice of properties codified in the annotation scheme. This led to the inclusion of literature which is primarily concerned with the organisation of discourse as well.

2.1 Anaphora processing and discourse

Halliday and Hasan [HH76] is the well-known seminal work which has inspired a large amount of research on cohesion in texts. The authors analyse in detail the relationships — named **cohesion ties** — existing between lexical items in an instance of discourse, both from a grammatical and a semantic point of view. The expanded concept of anaphora used in a great deal of subsequent research is in many ways similar to the notion of cohesion tie introduced in this work. The importance of chains of reference was also demonstrated within this context of textual cohesion, showing how the repeated reference to a certain entity, by means of a variety of linguistic devices, contributed to the organisation of a text. Conversely, the phenomena of pronominalisation and ellipsis could be understood satisfactorily when approached with textual aspects in mind.

Halliday and Hasan divide cohesion ties into five classes, namely: conjunction, reference, substitution, ellipsis, and lexical cohesion. Of those, conjunction is the only one not included in the expanded concept of anaphora mentioned above. The lexical items covered by the category — such as *however*, *on the other hand* and *notwithstanding* — signal semantic relations between clauses or sentences they connect. These relations are certainly an integral part of the way texts are organised, but they cannot be adequately characterised as anaphoric relations. The dependency on the identification of an antecedent for semantic interpretation does not describe the function of these words well. There is of course a degree of fuzziness in the boundaries between the classes which is explicitly acknowledged by the authors, but it seems adequate to set conjunction apart from the others.

Hobbs [Hob86] [1] describes two algorithms to resolve pronoun references in an NLP system. One is qualified as 'naive', as it relies on simple techniques which do not involve a large extent of semantic or world knowledge. The other is a more sophisticated attempt to link up all required information for anaphora processing, but, differently from the first one, it cannot be easily or at

[1] It is in fact an earlier paper, but it has been reprinted in the year quoted.

all implemented. The 'naive' algorithm works basically on syntactic surface information provided by parse trees, searching those trees to find noun phrases of the correct gender and number. It also incorporates the command relation as proposed by Langacker [Lan69] and simple selectional constraints on the noun phrases chosen. The first noun phrase found which satisfies the constraints is the antecedent.

The statistical results show that the 'naive' algorithm performs quite well, in fact better than any other semantically-based algorithm, because it does not require much processing sophistication. However, the only anaphors discussed by Hobbs are third-person pronouns when used to refer to explicit antecedents. Moreover, all the examples analysed by Hobbs are taken from written language. As it will be seen later, the 'naive' algorithm does not achieve such good results in the spoken language material analysed in this research. However, the first-candidate strategy for the resolution of anaphoric relations plays an important role in the assessment of recency as a factor in the processing.

Webber [Web79] focuses on discourse anaphora, thus widening the scope of phenomena included in the investigation. The listing presented in the introduction of the work shows the broad range of phenomena which the term discourse anaphora is meant to cover. The author concentrates on three types of anaphora, namely, definite anaphora (which includes pronouns and nonpronominal anaphoric noun phrases), one-anaphora, and verb-phrase deletion. The variety of types of knowledge required for the interpretation of anaphoric relations is focused upon, stressing the need for information which is not 'purely linguistic'. This refers to a conceptual model of the discourse built by participants.

The problem of identifying what an instance of discourse makes available for anaphoric reference is strongly emphasised, shifting away from previous work on anaphora. Before, the approach was primarily directed to the specification of constraints on the possible set of candidates for correct antecedents of a given anaphor, till only one choice was left. The notion of 'current context' is seen as an important way to constrain this set of possible antecedents, but the problem of identifying the specific elements which are part of this context is highlighted. This raises the question of the relationship between topicality and anaphoric phenomena, setting elements in salient topical roles as preferred antecedents for anaphoric expressions.

Sampson [Sam87], in a paper concerned with machine translation, points out the risks of dealing with problems of natural language processing on the basis of invented examples. The fallibility of scientific theories about observed facts — language included — is underestimated because the open-ended nature of the problem is not fully taken into account. As a result, proposed solutions treat natural language as if it could be reduced to a purely deductive problem. This amounts to a misinterpretation of the way human inferencing works. The author discusses a number of occurrences of the pronoun *it* extracted from the Lancaster-Oslo-Bergen Corpus, showing how the identification of a referent often depends on an ability to ignore ambiguities, rather than in algorithms to resolve them.

Therefore, the collection of a reasonably representative number of occurrences in real-life situations is likely to offer useful insight into the problems of natural language, and particularly into the resolution of anaphors. The paper also draws attention to the problems in the conclusions in Hobbs [Hob86], quoted above, where the 'naive' algorithm is seen as inferior to the hypothetic semantic one described later. Hobbs seems not to be satisfied with an algorithm that does not use inferencing in a deductive way, in spite of the acceptably good results. Sampson argues that 'tricks that work more often than not' are a much more feasible goal in natural language processing. Good results may be achieved by methods which bear no relation to the way humans would process language, given that one could actually be sure about the strategies of human processing.

Fox [Fox87] investigates anaphora on the basis of its relationship with discourse structure, also building on the notion that the possible set of antecedents can be constrained by an understanding of focus[2]. Fox's study is restricted to third-person singular pronouns which refer to humans.

[2]The confusion of terminology involving words like **theme**, **topic**, **focus**, **centre** and other similar expressions is

The analysis is developed on the basis of a sample which includes spoken and written language extracted from real-life conversations and articles in newspapers and magazines. Regarding the spoken-language data, Fox uses conversational analysis [SSJ74] to define the hierarchical structure of discourse, which is seen as crucial for the investigation. In accordance to this approach, the adjacency pair is the basic unit used to identify the structural organisation of discourse. Fox discusses the occurrences of long-distance anaphora found in the sample through this basic unit of conversational analysis, showing that such occurrences are possible when an adjacency pair is tied to another adjacency pair which is not the immediately preceding one. This kind of structure is called a **return pop**, a name borrowed from grammars based on augmented transition networks.

Pronouns would then refer to antecedents in the 'tied' adjacency pair, and certain linguistic devices — such as lexical repetition — would be used to 'mark' this return to a noncontiguous adjacency pair. A problem of circularity arises here, as the pronoun is said to signal that the pair is a return to a previous unit, but at the same time it seems to be necessary to identify this new unit in order to link the pronoun to a distant referent. In some cases, there are acceptable alternative antecedents between the anaphor and the correct antecedent. It seems important, consequently, to explain the processing strategy which allows the felicitous resolution of such long-distance references. The author acknowledges the circularity problem, but pursues the analysis without attempting to solve it.

2.2 Discourse models

Grosz and Sidner [GS86] extensively discuss discourse segments as units by means of which the structure of discourse is defined. The discourse structure 'is a composite of three interacting constituents: a linguistic structure, an intentional structure, and an attentional state.' Utterances are the basic elements of the linguistic structure, aggregating 'naturally' in discourse segments. The boundaries between segments may be controversial, but, according to the authors, different people will carry out the segmentation with roughly similar results. The intentional structure distinguishes a discourse purpose, which is the overall intention for the participants to engage in a particular discourse, and discourse segment purposes, which are intentions to be fulfilled by each segment. The attentional state 'records the objects, properties and relations that are salient at each point in the discourse', a process that is modelled by means of a stack, where focus spaces are associated to each segment and pushed onto or popped out of the stack as the sequence of segments evolves.

The attentional state constrains the interpretation of referring expressions, by limiting the focus space associated to each segment. The constraints on the use of pronouns within a segment are different from the ones which apply across segment boundaries. Discourse segmentation is thus one of the factors which governs the use of referring expressions. The theory is used to analyse two examples of discourse, one from a written argument and another from a task-oriented dialogue. The intentional structure is particularly complex. It is elaborated by means of distinctions between structural relations, such as **dominance** and **satisfaction-precedence**, which are difficult to recognise in authentic language. The idea of discourse segmentation, where each segment is associated to a certain set of elements in focus, is nonetheless a very important one for the study of anaphoric relations.

Grosz et al. [GJW95] present a theory for modelling local coherence within a discourse segment — i.e., coherence among utterances in that segment — referred to as the centreing theory. The theory is a development of the theory of discourse structure spelled out in the paper mentioned above, which aims at global coherence, that is, coherence holding among segments. Centres are then those elements in an utterance which link it to other utterances in the discourse segment. They are not a property of sentences, but of utterances in a discourse. Therefore, the same sen-

well known and to a certain extent inevitable. The author herself comments on the problem and points out the looseness with which they are used.

tence can have a different centre within a distinct context of discourse structure. Each utterance in a discourse segment is assigned a set of forward-looking centres. These forward-looking centres are ranked according to relative prominence. Each utterance other than the initial utterance in the segment is assigned a single backward-looking centre.

The elements which realise the forward-looking centres in an utterance are more likely to be the backward-looking centre of the following utterance according to their prominence ranking, as the backward-looking centre is the most highly ranked element which is realised in the following utterance. Grosz et al. use these links between centres to derive rules for the realisation of centres as pronouns or noun phrases. One important rule for the study of anaphora states that no element in an utterance can be a pronoun unless the backward-looking centre is also realised as a pronoun. This rule constrains the choice of referring expressions, and local coherence is compromised when it is violated. The examples given, however, are single-speaker texts and do not seem to have been extracted from real-life conversations, as they are very neat.

Hoey [Hoe91] explores cohesion, which the author deems an objective feature of text organisation,[3] by means of repetition-replacement relations. Coherence, on the other hand, is 'in the eye of the beholder', and thus subjective. Notwithstanding, the presence of such lexical patterns, as an essential aspect of cohesion ties (in the Hallidayan sense), contributes significantly to coherence. These patterns form interacting chains in order to produce text organisation, as the cohesive effect holds for long stretches of text. Therefore, it is possible to draw coherent abridgements out of a text by observing the patterns of cohesion and selecting those sentences where the patterns show most, that is, where the connections created by repetition and replacement are more numerous, no matter how far apart these sentences are. It should also be possible to track the development of topics in the text, by bringing together the sentences that share lexis, as well as to define the beginning and end of those topics 'in a principled way'.

Hoey then carries on to develop an approach that goes beyond the simple counting of repetition-replacement relations and involves matrices showing the connections between sentences in their full complexity. The final result is a system to analyse text and produce abridgements based on the analysis of cohesion processes. Hoey is careful to note that the system only works in non-narrative written texts. It is unlikely that patterns of lexis will show so clearly in the less orderly reality of spoken language. However, the notion that cohesion in text can be studied, at least to an important extent, by analysing lexical patterns in text seems to be sound enough to be explored also in spoken language. It is equally true, though, that any actual system, such as the matrices developed by the author, will certainly require extensive adaptation for use in spoken language.

Sinclair and Coulthard [SC92] propose a descriptive model for spoken discourse, based on a rank scale which comprises five levels: act, move, exchange, transaction and interaction. The lowest level is the act, which is roughly equivalent to the clause in grammar. An act is a functional unit, however, and thus different from a clause in important ways. The three major acts — elicitation, directive and informative acts — probably occur in all forms of spoken discourse. In their analysis of classroom speech using the model, Sinclair and Coulthard list several other types of act. Acts combine to form moves, which are the next level in the scale. Roughly, the act would be associated to an utterance, or tone unit [FH92], while the move would be the full stretch of speech produced by one participant before the other participant takes over.

The exchange structure is made up by a basic sequence of moves, namely: initiation, followed by a response, with a possible follow-up move. Only the initiation move is essential for the existence of an exchange. Those two or three moves combine to form an exchange, which contains contributions by two participants. A number of exchanges combine to form a transaction. However, the work was produced on the basis of data extracted from classroom interactions, where the three-move structure is the norm, with the third move optional. The boundary moves which indicate the end of one exchange and the beginning of another may be difficult to recognise in

[3] Hoey prefers "text organisation" to "text structure", as he believes the second is too strong a phrase to represent the reality of text. The point will be returned to in chapter 3.

other spoken-language situations.

Francis and Huston [FH92] use Sinclair and Coulthard's model to analyse a complete telephone conversation between two native speakers of English who are close friends. The problems with defining moves in a way as to make clear the exchange structure, and thus the boundaries between exchanges, become more apparent in the distinct context. The authors discuss the complications with the second move within the structure of the exchange, which involve the notion of predictability, or obligatory character. They conclude by stating that the minimum number of moves in an exchange is two, but are forced to consider a silence as a response move in certain circumstances. They are also forced to classify certain exchanges as incomplete. As expected, it is extremely difficult to find an analytic system able to cover all possibilities in spoken discourse.

Nonetheless, exchanges combine to form a transaction, which Francis and Huston characterise as basically a topic-unit, although they choose not to go into the 'thorny question' of defining topic. According to them, topic 'must remain a pre-theoretical and intuitive notion'. The authors state that the transaction boundary is recognisable by means of the intonation contour, but they add that the presence of such a contour is a necessary but not sufficient condition for the boundary to be established. They also acknowledge that it is impossible to define accurately a combination of exchanges which cannot happen. Therefore, the transaction is still a less satisfactory unit, as compared to the lower ranks of the scale. Without a definition of topic, it seems also difficult to be sure about the exchange boundaries.

Sinclair [Sin92] proposes two basic mechanisms of coherence — namely, prospection and encapsulation — which are at the foundations of the exchange structure. The first one results from a initiation move **I** which binds the interpretation of the following one, as it 'prospects' that the response **R** 'will be interpreted under the same set of presuppositions as the initiation itself'. The second mechanism is called encapsulation. It is associated to the 'third move', that is, a follow-up move **F** in an exchange which 'contains a reference' to the initiation-response or prospection-prospected pair. This basic mechanism is made more sophisticated in the analysis of written discourse in Sinclair [Sin93], where prospection is characterised as the typical topic-selection sentence, followed by a prospected sentence. Encapsulation classifies the sentences which refer to and constantly incorporate what has been said before. This is accomplished by means of logical acts — such as connectors — or deictic acts — such as demonstratives - which refer to previous discourse.

One example[4] of the basic three-move structure is shown below:

(4) **I** Why? Did you wake up late today ?
R Yeah, pretty late
F Oh dear

The question form is typical of the initiation move, resulting in prospection-type coherence, although responses may appear in question form, as in the following exchange[5]:

(5) **I** It's red
R Dark red?
F Yes

However, a challenge move **C** may interrupt the sequence, after either initiation or response moves, starting a new exchange. Thus, a challenge move is a form of initiation move which follows another initiation move, breaking the presuppositions for a prospected response. In the example below[6], the question introduced by the main clause *what do you mean* refers to the language used by the first participant and is not directly about getting up.

[4]The dialogue is part of the data analysed by Francis and Hutton [FH92], but it is also quoted in Sinclair [Sin92]
[5]Ditto.
[6]Ditto.

(6) I I was supposed to get up at about seven o'clock

C What do you mean you were supposed to

Sinclair [Sin92] acknowledges the problems involved in the single coding of moves. Given the complexity of human behaviour, assigning a single classification out of a small set of choices is bound to fail to describe its full effect. However, the author rightly points out that alternatives to single coding are not much better, as an exhaustive description is impossible, and a selective description 'invidious'.

2.3 Corpus-based approaches

Biber [Bib92] analyses the distribution of anaphoric expressions in 58 texts taken from nine spoken and written genres. One of these genres is face-to-face conversation. His research combines this corpus-based comparative approach with automated computational analysis of the texts in order to handle large quantities of data. In addition to the analysis above mentioned, Biber's paper intends to demonstrate the usefulness of this combination. The distribution of anaphoric expressions is assessed according to a set of features representing: the overall occurrence of given and new information; the overall frequency of referential expressions; the number and length of anaphoric chains; the distance between referring expressions within chains; and, for given or anaphoric expressions, the choice between lexical repetition and pronominal forms, as well as the syntactic distribution of forms (subordinate or main clause).

These features are classified according to a distinct set of categories for each one of them. A statistical procedure called the General Linear Model is used to analyse the frequency counts obtained. Concerning conversation, Biber's results show that, although the genre shows one of the highest frequencies of referring expressions, it has the lowest frequency of different referents. Referential chains constitute a relatively high proportion of the different referents in conversations, that is, referents which are mentioned only once are less frequent in the genre. The chains in conversation are the longest by far, where chain length is measured by the number of referring expressions included in a chain. These measures may be used to characterise conversation as a genre where a small number of different topics are dealt with, but they are discussed at length. Many other interesting findings are discussed in Biber's paper, but they cannot be fully accounted for here for reasons of space.

Finally, Fligelstone [Fli92] developed an annotation scheme to analyse anaphoric relations in corpus material, the only work found which is truly similar to the present research in that sense. Fligelstone's motivation is also to investigate the possibility of building a probabilistic device to resolve anaphors. In order to develop anything of the kind, the need for a large amount of pre-analysed data is acknowledged. Annotated corpus material thus contributes for the creation of a database of this sort. The material would also be useful for research concerning anaphora in general, as data can be extracted from the database for such purposes. Fligelstone points out that the adoption of a 'fairly coarse-grained scheme' has the advantage of being, to a certain extent, theory-neutral.

The scheme uses numbered angle brackets to mark any section of text considered to be an anaphor. The antecedent is also marked in this way, and the numbering shows the tie between the anaphor and the correct antecedent. In addition, the anaphor is classified according to: direction, which is basically anaphoric or cataphoric, indicated by the bracket pointing left or right; type of relationship, which includes some of the concepts introduced by Hallyday and Hasan [HH76] as forms of cohesion ties, plus a few other possibilities; the identification of the antecedent by number, with some options to indicate multiple or uncertain antecedents; and additional features, encompassing a variety of symbols to codify grammatical number, secondary reference, degree of uncertainty and other phenomena.

2.4 Summary

Fligelstone's work is the one which bears most similarities to the research described in this paper, as it also develops a corpus annotation to analyse anaphoric relations. There is a lot in common also with Biber's methodology of using statistical procedures to quantify features of discourse relations. This project believes statistics can be used to predict features of the antecedent, once features of the anaphor are given. Sampson's concern with the use of authentic data is fully shared and clearly adopted as a principle here, as previously stated. The discussion on processing in Sampson is also a crucial matter which this research will focus upon intensively.

Hobbs' naive algorithm is the starting point, as it were, for this investigation on processing. Anaphor resolution would be greatly simplified if the algorithm could be successful in finding the correct antecedent in all instances. It is known beforehand that it won't be, but the research sees the recording of a processing strategy for the resolution of all cases of anaphora found in the corpus as essential. Therefore, the extent to which the naive approach is successful in dealing with anaphors in spoken language will be assessed and an alternative processing strategy proposed whenever it is ascertained that the correct antecedent would not be found by such a means. An estimate of probabilities for the success of given processing strategies according to type of anaphor should be possible as well.

This brings up the question of topicality and discourse organisation as ways of limiting the search space. It seems to be generally agreed that salient elements in discourse are preferred antecedents. Therefore, it is intuitively plausible to expect that processing strategies involve the use of discourse information to keep track of salient elements, possibly by segmenting discourse in such a way as to restrict the list of available candidates at any given point. Thus, Grosz and Sidner's work on segmentation is important for this research, as well as Grosz et al. as an attempt to model coherence. However, the ideas presented therein are adopted in their broad lines only, not with respect to specific methods of segmenting discourse and selecting the elements in focus.

This is due, inter alia, to the sort of inferencing involved in dealing with Grosz and Sidner's intentional states, which requires modelling that clashes with the empirical methods preferred in this research. The experience of analysing corpus data showed that the form of discourse organisation proposed is not easy to identify in authentic spoken language. The approach to cohesion in Halliday and Hasan, further developed in Hoey, proved to be more flexible and adequate for the requirements of this research at a global level of discourse analysis. The formulations in Sinclair, Sinclair and Coulthard, and Francis and Hunston were important for the segmentation methods at the level of local coherence and identification of segment boundaries.

If segmentation and topic tracking are to play a truly useful role in the analysis of anaphoric relations, as they seem to have the potential to, it is essential to find a way to account for these phenomena that is straightforward enough for practical use by analysts dealing with corpus material. It is also important to overcome the circularity risk pointed out by Fox. This means that the segmentation must be carried out independently of anaphora interpretation. It is expected that segmentation will help identifying the referents of anaphors. Therefore, the identification of segment boundaries and topic continuity cannot rely on the resolution of anaphoric references. The next chapter details the problems encountered in integrating the notions of anaphora, topic, and discourse segmentation into a coherent approach, as the analysis of corpus data progressed. The discussion concludes with a rationale for the features of the annotation as it currently stands.

Chapter 3

Methodology

Attempts to analyse anaphoric relations in samples from the London Lund Corpus showed that the scheme presented in [Fli92] was not adequate for the demands of spoken language. The strategies required for the resolution of anaphors which could not be handled by recency techniques involved segmentation and topicality. Moreover, the separation of types of anaphors and processing strategies as two distinct properties handled the cross-linguistic contrastive analysis much more effectively. As the scheme presented in [Fli92] was not meant to analyse spoken language nor for cross-linguistic analysis, it is of course not surprising that it could not be easily extended for these purposes.

The inclusion of topicality and processing as properties to be analysed together with a classification of anaphors and antecedents could not be accomplished within the specifications of Fligelstone's scheme. Although some elements of processing were scattered between two variables, the variable called **type of relationship** also involved the type of anaphor used without a clear-cut distinction. It was thought best sharply to separate the classification of the anaphor from the processing involved in resolving it, for reasons discussed in 3.2. Soon it became clear that an annotation scheme would have to be created specifically for the purposes of the current research.

The envisaged annotation scheme contains three primary sources of methodological difficulties, namely: the notion of topic; problems in carrying out discourse segmentation; and complexities involving the properties singled out in the classification of anaphoric phenomena. The first two are strongly interrelated, whereas the third one is, to a certain extent, independent. All of them have important bearings on the features of the annotation. The discussion of methodological aspects begins with the pitfalls on the way towards an operational definition of topic.

3.1 The notion of topic

Intuitively, task-oriented and information-seeking dialogues, which are the sort of conversation with which this study is concerned, are 'about' something which participants implicitly or explicitly agree is the topic of their conversation. However, if asked, two participants in a dialogue may use distinct phrases to sum up the gist of their conversation. This does not mean one of them is wrong. If dialogue S02.01.01 in the RLLC were classified as about *financial problems*, this could not be said to be a mistake. Nonetheless, *the funding of Mr. B's bibliography* would be a perfectly reasonable and perhaps better account of the dialogue topic. Several variants on the same general idea could come up as other people who read or listened to the dialogue were consulted, ranging from a laconic *finance* to a prolix *the problem Mr. B had when he went to see Mr. A in order to ask for support in his attempt...* and whatever might follow.

The first problem the analysis was faced with was to formulate a definition of topic which could be used for the purposes of the annotation scheme. A clear measure for the difficulty of

the task is the well-known confusion involving the topic-related terminology (see [BY83] and [Fox87]). Nonetheless, for the purposes of this research, it was most important to define topic in a way that different analysts could use the definition and reach similar conclusions, or else the annotation scheme would have to abandon all claims of usefulness. The reliability of statistical results would be jeopardised as well, because there would be variables dealing with a concept of a vague nature which could not be used without unpredictable fluctuations.

Moreover, the purpose of incorporating topicality into the annotation scheme is to support anaphora resolution. It is believed that a significant share of those cases of anaphora which cannot be handled by a Hobbs-like algorithm would be successfully handled by an anaphora interpreter if topic tracking were improved. This belief is the rationale for the resolve to find ways to include the notion of topicality in the analysis of anaphoric phenomena. Therefore, the definition of topic sought is one that is most conducive to the purposes of anaphora resolution, not necessarily the one which produces results closest to intuitive judgements, even if they were assumed to be uniform. Intuitive plausibility is clearly desirable and taken into account, but feasibility of mapping into a procedure, as well as adequacy to the purposes of anaphora resolution, are more important in the context of the annotation scheme.

Given these assumptions, the prolix version of the topic mentioned above would be poor help for the purpose of anaphora resolution, as it does not stand a very good chance of being referred to. The observation of anaphora cases showed that generics such as *finance* or *funding* — and, even worse, *problem* — are too abstract to be frequently referred to. The actual references made in the text are more likely to be specific forms of *finance*, such as *Ford Foundation grant* or *Canadian money*. Selecting a generic often means that the topic chosen will not be a salient element as far as frequency of reference is concerned. One way around the problem would be to group several elements such as these under *finance* and call them all instances of the global topic for the dialogue. However, that is likely to generate a lot of dissent among different analysts as to which elements to include. The grouping of various elements under an umbrella topic would also cause the number of discourse entities permanently held as salient to grow unnecessarily, requiring ways of choosing between them for the processing of individual anaphoric references.

The second problem relates to the distinct levels of topicality contained in an instance of discourse, such as a dialogue. It is easy to notice that a global topic in a dialogue typically branches into local topics which prevail as the element of highest saliency in a stretch of discourse. They are related but not identical to the main global topic. Moreover, the relationship between a global topic and the local topics is not necessarily obvious from a semantic point of view. On the contrary, it is often built within the context of a specific dialogue. No semantic net would consider the nouns *will* and *tray* to be connected in any plausible way. Nonetheless, they are related as global and local topic, respectively, within one of the dialogues analysed.

Local topics are intensively referred to throughout the stretch in which they are the dominant element. It is clearly important to include them in an account of topicality. Moreover, many dialogues are complex enough to require further division of segments into subsegments, each one of them with its own subsegment topic. It is also possible that a given dialogue contains a radical change of subject matter to an entirely unrelated topic. Both situations have to be handled in some way by a topic-tracking mechanism which would be possibly developed on the basis of the proposed topicality account. This does not mean to say that dialogues where there is one single topic and no room for any local topics never occur. Short conversations strictly aimed at obtaining one single piece of information may require no further elaboration of the notion of topic. However, this is not the sort of dialogue included in the sample analysed in this research.

Other entities may have salient positions in the topical hierarchy of a dialogue without being topics. It has been observed in previous works on topicality (see [vDK83], [GS86], [All87]) that references are made to elements which are related to global or local topics. The elements with a global presence which are not the discourse topic may become local topics at certain passages. Others are agents or participants in the dialogue, who play an important role throughout without

ever being the dominant topic. These elements related to the global topic should also be specified together with the global topic for the dialogue. The same holds for each segment. A chosen topic should be specified, along with a set of salient elements, for each segment. The analysis of topicality contained in the annotation scheme must account for these ancillary elements in a satisfactory way.

The observation of corpus data does seem to offer support for the focusing spaces attached to segments proposed in Grosz and Sidner [GS86]. The approach used here simplifies these focusing spaces by disconsidering notions such as purpose and intent. Focusing spaces contain solely lexical items found in the dialogue. It is useful to select the elements included in such focusing spaces on the basis of global (or discourse) topicality and local (or segment) topicality. The number of elements with a status of global topicality, which need to be permanently available in focusing spaces throughout the dialogue, appears to be small in spoken language. Therefore, there is no need to retain a full history of entities in past segments along with each focusing space. In fact, these elements of global presence can be tracked by using an adaptation of the approach created by Hoey [Hoe91], relying primarily on lexical cohesion. Those items of local saliency would be added and then deleted as the dialogue progresses, while the ones of global significance would be invariably available for reference.

Hoey's system involves the use of repetition-replacement relations to establish lexical patterns that show cohesion. This includes the resolution of anaphoric references, which is one of the forms of cohesion tie, as defined by Halliday and Hasan [HH76]. In fact, it is often difficult to draw a line between the two concepts. However, as previously pointed out, it is essential to avoid circularity. As the research assumes that topic tracking is a crucial aspect of anaphora resolution, it cannot rely on the resolution of pronouns to track the topic. It was decided then that the only forms of cohesion to be included in the procedure would be simple lexical repetition (*bibliography/bibliography*) and complex lexical repetition (*bibliography/bibliographical*), links which could be established with plain search mechanisms.

The initial procedure for topic tracking used for the annotation work relied strictly on the analyst's intuitive decision as to what was being talked about, with all the problems that have been mentioned before. This was done by simply reading the dialogue in full and then reaching a decision as to the best phrase — typically a noun phrase — to sum up the gist of the interaction. The decision was made before the anaphora cases were individually annotated, keeping choices related to topical structure separate from those related to anaphoric relations. Not only the discourse topic, but the whole set of segment and subsegment topics was fully worked out before the annotation of anaphors began. The procedure sought should not alter this basic approach, even when using cohesion ties to make it more objective and replicable.

In order to introduce the topicality component in the analytical scheme, a topic — called the **discourse topic** — was assigned to each dialogue analysed. In case it contained a radical change of topic, the dialogue was split into fragments, each one with a separate discourse topic. The dialogue or dialogue fragment was then divided into segments, and each segment was assigned a topic — called the **segment topic**. Consequently, a new segment is started when a local topic shift occurs. It is not unusual that certain segment topics are further developed into subordinate topics, retaining nevertheless its saliency throughout a relatively long stretch of discourse. Such stretches are divided into subsegments, each one with a topic — called the **subsegment topic**. A new subsegment is started when a subordinate topic shift occurs. A set of highly prominent elements associated with the discourse topic were selected under the name of **discourse thematic elements**. A similar set was chosen for each segment, called the **segment thematic elements.**

Segment and subsegment topics may resume in a dialogue which revolves around the same discourse topic. As the discourse segmentation proposed is based on topic continuity, the resumption of segments and subsegments means the participants in a dialogue return to a topic previously developed in a segment or subsegment which is not the immediately preceding one. An identical or very similar set of salient entities is recalled together with a resumptive segment or subseg-

ment. There are several distinct possibilities of topic resumption. A segment topic which had not been developed into subordinate topics may reappear by returning to the segment topic itself or by starting a subsegment with a recognisably subordinate topic[1]. A segment topic which had been developed into subsegment topics may reappear by returning to the segment topic proper; by returning to a subsegment topic previously developed; or by returning to a new subsegment topic which is recognisably subordinate to the segment topic in question. A subsegment topic may resume within the segment or, as just mentioned, as a discourse unit within a resumptive segment topic. The annotation scheme attempts to codify all those variations for the purposes of topic tracking and the resulting segmentation.

The sample for the English language contains six dialogues of different lengths. All of them required division into segments and subsegments in order to adequately represent topic development and the link between topicality and reference. Only two dialogues needed breaking up into fragments with radically distinct discourse topics and separate topical structure. The shortest dialogue in the English sample contains 854 words and the longest 7741 words. The division of segments into subsegments is likely to be necessary even in shorter dialogues, but it is not difficult to think of interactions without subsegments and, as said before, even without segments. They would have to be very short, though. On the other hand, the separation into fragments is unlikely to be associated with length in any way. It is the radical change in subject matter that makes the difference.

The following subsections define each one of the topical roles in the topicality hierarchy and specify a procedure for their identification. It was imperative to reach tractable definitions for the different topical roles that could be mapped into straightforward identification procedures. By 'tractable' is meant definitions that by and large lead different people to the same conclusion concerning the elements in the topical roles. It should be kept in mind, as explained above, that the intuitively satisfactory choice may not be the one that serves the research purposes best. In other words, a definition of topic which can be successfully mapped into a straightforward procedure is preferable to one which satisfies general intuitions but does not foster consensus.

3.1.1 The identification of a discourse topic

The attempt fully to formalise a notion such as 'topic' is doomed to failure. Nevertheless, the limits to formalisation were explored to push them as far as possible, although the procedures to identify the topical roles ultimately require a decision by a human analyst which cannot be formalised. This required a process of analysing, testing and comparing the different dialogues until a practicable procedure was settled upon. This process is by and large described in the following pages.

Four features are taken into account in the selection of the best candidate for discourse topic, namely, frequency, even distribution, position of first token, and semantic adequacy. The discourse topic must be a lexical item which is frequently referred to. As discussed above, generics may be intuitively attractive, but there are dialogues in which an intuitively satisfactory generic is seldom referred to throughout the discourse. Assuming a machine-readable POS-tagged dialogue, the first step would be then to run a word frequency count to determine which words appear most frequently in the dialogue. The most common words in an instance of discourse are grammatical words, such as pronouns, articles, prepositions, and conjunctions. These are not suitable for the function of discourse topic and can be eliminated as candidates. Therefore, the feature of semantic adequacy is already playing a role at this point. Certain noun phrases of unspecific semantic content, such as *thing*, *sort*, and *fact*, can also be eliminated.

Returning then to dialogue S02.01.01 (hence Dialogue 1) in the RLLC as an example, the top of the frequency count list for suitable lexical items appears as below. The figures shown conflate singular and plural tokens of the lexical items. Occurrences resulting from repetitions in hesitations and false starts have been screened out. Given a POS-tagged corpus, these adjustments

[1] The problem of recognising subsegment topics will be addressed below.

should be possible by automatic means.

bibliography	=	16
Ford	=	14
press	=	14
university	=	14
English	=	13

Considering the small differences between the frequency totals for each lexical item, a decision cannot be made without bringing in new criteria. The notion of even distribution is then brought to bear on the frequency data. The distribution of the lexical item throughout the dialogue is a crucial factor for the identification of the topic, because a high-frequency item could occur many times in a relatively small stretch of the dialogue and then not occur any more. This lexical item would be unlikely to be the discourse topic, which is expected to be more evenly distributed. In the dialogue presently used as an example, there are 1160 tone units — henceforth referred to as lines. The first and last lines in which each one of the lexical items are shown below.

bibliography	0047-0891
Ford	0067-1129
press	0130-1004
university	0232-0998
English	0073-0924

The information on first and last lines does not seem to make things clearer in any obvious way. Occurrences spread in a reasonably even manner throughout the dialogue, although none of the high-frequency items selected occur right at the beginning. One way to make things more precise is to calculate means of incidence for the whole dialogue, relating frequency to number of lines, and compare them to means of incidence for the stretch between the first and last lines in which the lexical item occurs. If the means are significantly different, the lexical items can be considered as not having an even distribution. A ratio can be calculated by dividing the mean for the whole dialogue by the mean for the stretch in which the lexical item occurs.

Numbers in the second column refer to the mean obtained by dividing the total number of lines in the dialogue by the total number of occurrences for each lexical item selected. The third column shows the means obtained by dividing the total number of lines between the first and the last occurrence of the lexical item by the total number of occurrences for each lexical item. The fourth column presents the ratio obtained by dividing the number in the second column by the number in the third column. Concerning the first lexical item, *bibliography* thus occurs roughly every seventy-two lines in the dialogue and every fifty-two lines in the stretch between the first and the last occurrence, with a distribution ratio of 1.37.

Table 3.1: Distribution of high frequency lexical items in Dialogue 1

	whole-dialogue mean	first-last mean	distribution ratio
bibliography	72.50	52.75	1.37
Ford	82.85	75.85	1.09
press	82.85	62.42	1.32
university	82.85	54.71	1.51
English	89.23	65.46	1.36

The data are inconclusive. The differences between the ratios for each lexical item are quite small. Moreover, none of these ratios is significantly high, that is, none of these lexical items can be said

to have an uneven distribution in the text on the basis of these results. In order to contrast these data with a dialogue where distribution clearly rules out a high-frequency candidate for discourse topic, compare the numbers for dialogue S11.01.00 (henceforth Dialogue 3). The dialogue has 1288 lines. Token totals conflate singular and plural occurrences. Figures are adjusted for repetitions in hesitations and homographs — such as the future auxiliary *will* — which cannot be considered as a possible discourse topic — and the noun phrase *will* — which is the discourse entity being considered for the topical role. Results are shown in 3.2 below.

Table 3.2: Distribution of high-frequency lexical items in Dialogue 3

	total of tokens	stretch of incidence	whole-dialogue mean	first-last mean	distribution ratio
mother	60	0009-1260	21.46	20.85	1.02
doctor	42	0160-0669	30.66	12.11	2.53
will	39	0006-1283	33.02	32.74	1.00
coleman	22	1103-1288	58.54	8.40	6.96
kay	21	0121-0953	61.33	39.61	1.54

Differences here are much greater. Thus, *coleman* — which is an extreme case of concentrated incidence — and *doctor* were eliminated on the basis of distribution, in spite of the high frequency scores, especially in the case of *doctor*. Another important component in the process of choosing a discourse topic is the position of the first token in the dialogue. It is expected that the discourse topic should be introduced early in the dialogue. It is not advisable to ignore such a strong intuitive factor in a procedure for the identification of the discourse topic. The fact that *doctor* appears for the first time in line 0160 would make it an unsuitable candidate for the function, one which compares unfavourably with *mother* (line 009) and *will* (line 006), the competing best candidates. The decision to select one of those two will be discussed later.

Returning then to Dialogue 1, *bibliography*, which has some preference over other candidates for being the most frequent lexical item, is the best choice regarding first-appearance position. *Ford* and *English* do not fare much worse, but the other two candidates appear for the first time late in the dialogue. On the other hand, the gap in the end of the dialogue is larger for *bibliography* than for *Ford*, although experience reveals that long gaps in the end are much more tolerable than those in the beginning. Once several elements are firmly established in the course of the dialogue, it is not uncommon to have one of them prevail as a topic for a stretch, even when it is not the discourse topic. This is not as likely in the beginning. In addition, tokens at the very beginning or end may be outliers, which warp the distribution ratio. Using a command such as `grep -n` in Unix^{TM}, the next step is then to look at the pattern of incidence more closely, so as to see what else can be learned from the distribution of high-frequency lexical items in the dialogue. The line-by-line map for each high-frequency lexical item in Dialogue 1 is shown below.

bibliography 047; 0253; 0256; 0260; 0262; 0269; 0336; 0341
0418; 0648; 0675; 0693; 0887; 0890; 0891
Ford 067; 082; 088; 0104; 0481(two); 0487; 0493
0584; 0618; 0643; 0755; 0818; 1129
press 0130; 0232; 0246; 0253; 0254; 0274; 0280
0344; 0359; 0362; 0368; 0777; 0992; 1004
university 0232; 0274; 0280; 0298; 0330; 0344; 0359
0362; 0368; 0433; 0777; 0956; 0994; 0998
English 0073; 0221; 0264; 0283; 0288; 0598; 0598
0600; 0602; 0784; 0865; 0911; 0924

One important new information about the distribution is the gap of more than three hundred lines

between the last token in *Ford* and the one before the last, which characterises an outlier. This adds to the preferential status already enjoyed by *bibliography* as a candidate for discourse topic. It is true that both *bibliography* and *Ford* show relatively large gaps — over one hundred lines — where there are no tokens. However, the longest gap for *bibliography* is between lines 0418 and 0648, a 230-line gap. *Ford* shows a 377-line longest gap between lines 0104 and 0481, which increases the prospects for the selection of *bibliography*. In addition, *Ford* also has a second very large gap already mentioned, which is the one between the two last tokens. This means the slightly lower ratio for the distribution of *Ford* does not measure evenness well. In fact, those two lexical items seem to take turns in prevailing over stretches of dialogue.

The balance tilts to *bibliography*, but it is still difficult to decide. One way of trying to find out some more about the patterns of occurrence for the words in question is to lemmatise the search (that is, to search for inflected and derived forms along with the basic word stem). The lemmatised search of *Ford* yields nothing at all, and that is not surprising. On the other hand, *bibliographical* appears three times in lines 0897, 0907 and 0910. In this case, it happens to be a rather significant finding because it extends the distribution range of *bibliography* by nineteen lines. Of course the adjective tokens do not count as if they were identical to the noun phrase, but the contextual analysis with `grep` shows that the tokens are truly linked to *bibliography* in a binding way.

The line-by-line map also reveals the coincidence of *university* and *press* in eight lines, suggesting the possibility of a noun phrase, the obvious one being *university press*. This would be automatically confirmed by the tags in a tagged corpus. The other six tokens of *university* would be tagged as noun-phrase heads, except for *University Microfilms*. This token is not a crystallised phrase such as *university press* and cannot be properly understood at this stage, as it requires contextual analysis for clarification. Nonetheless, this rules out *university* as too infrequent and sparsely distributed. The remaining six tokens of *press* are also noun-phrase heads. Although the qualifier *university* is not realised in these tokens, there is no reason to believe that they do not refer to *university presses* all the same. On the other hand, selecting *university presses* as a discourse topic would mean adopting a topic with a first token in line 0130 and a gap of over four hundred lines between lines 0368-0777.

The next step is then contextual analysis, which should be kept within as short a range as possible, typically two or three lines before or after the line of occurrence. This is the first moment when the analyst is supposed to look into the dialogue proper. The pattern of occurrence is checked in terms of consistency. For instance, *Ford* appears fourteen times according to the frequency count program, but it is not known whether it appears five times as *Ford Motors*, three as *my old uncle Ford*, and the remaining seven as *Betty Ford*, although it may be hard to figure out how such a combination of referents would occur in a single sample of dialogue. These tokens would have different referents and could not then be added up as tied in terms of cohesion for the purposes of topicality in discourse. Contextual analysis confirms that both *bibliography* and *Ford* refer invariably to the bibliography compiled by one of the dialogue participants, and to Ford Foundation, respectively.

The results of the contextual analysis seem to disqualify *English*. It appears in a variety of contexts which are difficult to relate, sometimes as a qualifier, others as a noun-phrase head with modifiers that point to different referents. There is no way to think of those occurrences as having a single referent. At least three times the lexical item clearly refers to the nationality and not to the language. Moreover, it appears later in the text than the other two strong candidates and it is the less frequent choice. It might be retained as a fallback option, but it does not seem to be the sort of element which is suitable to the function. Of course the actual reading of the dialogue in full will easily demonstrate that *English* is not the adequate discourse topic for Dialogue 1.

The six tokens of *press* indeed refer to university presses, but the whole set of tokens includes references to various university presses without reiteration of the qualifier. There are two generic references as well. These publishers are referred to as possible sources of funds for the bibliography at issue. They are therefore on a par with *Ford* in the context of this dialogue. Tokens

clearly concentrate on the large gap between 0104-0481 where *Ford* does not occur. The 0992 occurrence, which also relates to two tokens in the remaining *university* occurrences (the *London* tokens; all the other *university* tokens are not referred to as university presses or funding agencies, but as job options), also fills one of the *Ford* gaps. This probably means that *Ford* enjoys a special status among the possible sources of funds, which the reading of the full dialogue will confirm. These conclusions require expanding the context analysed beyond the typical `grep` line, but not unreasonably so.

All things considered, *bibliography* seems to be the most appropriate choice as a working hypothesis for the annotation work to start. The choice of a generic 'sources of funds' topic would have the advantage of being a more frequent topic if all sources are added, but, in fact, referents are distinct. Moreover, the common element which justifies the inclusion of all these funding agencies in the dialogue is the object to be funded, that is, *bibliography*. One possible solution would be to name *funding of the bibliography* as the discourse topic, but that brings back the problem of low-frequency reference or indirect reference in ways that are likely to undermine consensus among analysts. Step 8 in the procedure below specifies the preference for discourse topics which are items explicitly referred to in the dialogue, rather than generics comprising distinct lexical items referred to in the dialogue.

To sum up, the working hypothesis for Dialogue 1 in terms of discourse topic identification is that *bibliography* is the discourse topic. There is a short stretch of dialogue in the beginning which may be a separate fragment with a distinct discourse topic. *Ford* or *Ford Foundation* is a discourse thematic element, that is, a recurring element related to the discourse topic which is likely to be the segment topic in more than one stretch of the dialogue, characterising a resumptive segment. *University press* is also likely to belong in the set of discourse thematic elements. By default, the two participants in the dialogue are also discourse thematic elements. They are never selected as discourse topics, as they are so rarely referred to anaphorically in a dialogue in which they participate.

Persons in general, even when a third party, are not a preferred discourse topic, although they may be on the top of the frequency count list and present an adequate distribution. Dialogues are seldom about a person in general, but about some aspect of their lives or something they have done or participated in. One example of this is Dialogue 3, where *mother* reaches 60 tokens and *doctor* 42, whereas the noun phrase *will* adds up to 39. Nonetheless, reading the dialogue makes it quite clear that the topic in question is the *will*, not the person *mother*, and certainly not the *doctor*, eliminated because of distribution, as shown above.

This seems to contradict the initial claim that the priority was to spell out a straightforward procedure to establish the topic. However, it is generally more useful to single out an inanimate object as discourse topic, a status of unique saliency, than a person. Pronoun references to persons are more restrictive than the neutral references to objects, leaving less room for ambiguity. The set of possible antecedents is generally smaller as well and there are no possibilities such as the sentential *it*. In this dialogue, *will* stands out as the most frequent inanimate object by far. Diluting this amid the thematic elements does not help the resolution of the 132 occurrences of *it* in the dialogue.

Moreover, priority is not meant to be absolute precedence. The procedure should not be made to override strong intuitive claims. In Dialogue 1, the first analytical reading selected *finance* as the topic. The word is hardly referred to at all and only indirectly. It is definitively not a good choice. The topic *bibliography* is perfectly plausible and is also referred to many times. In Dialogue 3, it is difficult not to choose *will* as the topic. The procedure is meant to eliminate possible misunderstandings of the concept of discourse topic, not to force counter-intuitive judgements on analysts. Given that the choices available are high-frequency elements with an even distribution and basically the same referent, it seems wise to prefer objects instead of people. Step 9 in the procedure specifies this preference.

For the purposes of this research, the discourse topic is defined as the element of highest

saliency in a dialogue, where saliency is a function of frequency, even distribution, position of first appearance, and semantic adequacy. Even distribution is crucial. A highly frequent element which occurs intensively in a passage of the dialogue but does not appear for long stretches is not likely to be a good choice for discourse topic. This is particularly true if the first appearance occurs a long way from the beginning. Semantic adequacy has to be assessed by the analyst, and is not as objective as frequency and distribution. Some guidelines, based on the analytical experience accumulated so far, are set to help reduce the degree of unpredictability. The resulting procedure is shown below.

1. Run a word frequency count for the dialogue
2. Select the five most frequent suitable items, discarding:
   - grammatical words such as pronouns, articles, prepositions and conjunctions
   - noun phrases of unspecific semantic content, such as *thing*, *sort* and *fact*
3. Check the distribution of these items throughout the dialogue, selecting the most evenly distributed for a working hypothesis
4. Check whether the same item is linked to different referents by means of short-range contextual analysis and prefer those which have a single referent throughout or irrelevant variation
5. Check whether the referent for the high-frequency items is referred to by different items and analyse the effect of this on frequency and distribution
6. Lemmatise the search and consider the significance of lemmatised-token frequencies for global frequencies and distribution
7. Check the position of the first appearance and prefer the one closest to the beginning of the dialogue
8. Prefer lexical items explicitly appearing in the dialogue to generic noun phrases covering several items
9. Prefer items referring to objects instead of items referring to people
10. Check manually by reading the full dialogue

The last item may be carried out along with the first stage of the annotation work, which involves the definition of fragments, segments and subsegments according to topicality. It should be possible also to identify discourse thematic elements in the process of selecting a discourse topic. Both the definition and the procedure draw on the work on lexical cohesion by Hoey [Hoe91]. The selection based on frequency and distribution ensures that the discourse topic will be a useful notion for anaphora resolution and reduces subjectivity in choices, although it is not possible to eliminate it entirely, as cohesion is a property of the text, but coherence is a result of the reader's evaluation of a text. Therefore, step 10 may result in a complete reversal of previous expectations, although this has not been experienced in the annotation process carried out in this study. The procedure is not meant to push analysts into counter-intuitive choices, but rather to avoid purely subjective decisions and the problems described above. The danger of circularity is also avoided by considering only simple and complex lexical repetition as measures of cohesion.

The procedure is likely to run into trouble in many cases, as coherence does not follow easily from patterns of lexical cohesion. The analyst should be able to handle these situations with a flexible understanding of the procedure. A last example will be discussed, as it also seems to contradict one of the strategies defined in the procedure. The frequency count for Dialogue 2 (S01.02.03 in LLC) yielded the results below, unsuitable candidates omitted.

faculty	=	8
council	=	8
committee	=	7
academic	=	6
university	=	4

The item *faculty* is the most frequent, together with *council*. It appears six times as a qualifier in the noun phrase *faculty board* or *board of faculty*, which mean the same in the dialogue, and twice as head of the noun phrase *faculty of arts*. As a single-referent item, thus, it occurs in fact six times. The item *council* occurs four times in *academic council*, twice in *extramural council* and twice in *collegiate council*. Therefore, *committee* rises to the position of most frequent item. The `grep` command shows that it appears four times as *senate committee* and three other times referring to other senate committees. The occurrences spread from 1216 to 1350[2], with no tokens from 1351 to 1463. This is a short dialogue, so that the absence of tokens for 112 lines is quite significant.

The problem of reference is quite subtle here, because that is precisely what the participants are talking about, i.e. the *terms of reference* of the senate committees. The analysis of *board of faculty* and the three different phrases which include *council* shows that all these noun phrases refer to different senate committees. Moreover, the distribution of these phrases is complementary, filling the absence left by *senate committee* in a highly concentrated way. After the first token in 1319, just before the last occurrence of *senate committee*, *board of faculty* or its equivalent appear seven times in forty-two lines. In similar fashion, *council* appears for the first time just before the last token of *board of faculty* and then repeatedly for six times in less than forty lines, referring to three different senate committees. This chain covers virtually the whole dialogue.

Step 8 in the procedure states that generics covering several different referents in a dialogue are not to be preferred as discourse topics. However, there is simply no other appropriate choice here, as virtually everything hinges around senate committees. Consequently, the discourse topic chosen for this dialogue is the plural noun phrase *senate committees*, which encompasses *faculty board* or *board of faculty*, the three different *council* referents and the *senate committee* or simply *committee* tokens. Different numbers may be assigned to each one of these different referents in the referent list, but the topical role of the antecedent can only be annotated as the discourse topic. This highlights the fact that Step 8 may conflict with Step 5.

Again, the primary aim of the procedure is to avoid discourse topics such as *the day Robert went sailing in the river nearby and the boat capsized*, which would serve no purpose at all, but might be intuitively reasonable. The procedure tends to force choices such as *boat* or *river*. Analysts are sure to be confronted with decisions which are far less clear-cut than that, but hopefully the procedure will help narrow down the number of candidates, striking a sensible balance between intuitive plausibility and the aims of research on anaphoric phenomena. The next subsections specify similar procedures for the identification of the remaining discourse units previously defined.

3.1.2 The identification of a fragment

The analysis of three features of lexical items suitable for the topical roles — frequency count, distribution and position of first appearance — revealed a number of facts about the dialogue which would be annotated subsequently. Once associated to the analysis of immediate contexts of occurrence, the analysis led to a working hypothesis concerning the discourse topic or topics. Other mechanisms, such as lemmatised searches and the investigation of the argument structure for frequent verbs, added support for the hypothesis. The analyst would have to read the dialogue and confirm the hypothesis, but, once mapped into a procedure, the analysis of these features is likely to reduce the chances of variation from analyst to analyst to a minimum. The possibility of

[2]The first line in the dialogue is numbered as 1214, due to organisational characteristics of the London-Lund Corpus.

selecting a discourse topic which would play a weak part in the processing of anaphoric relations in a dialogue is virtually eliminated.

The analysis also allows the identification of a possible division into fragments with distinct discourse topics, by observing the distribution and the position of the first appearance for an item selected as a working hypothesis for discourse topic. The absence of discourse topic tokens for long stretches at the beginning and at the end of a dialogue should be analysed with the possibility of a fragment in mind. Gaps within the dialogue should in principle be seen as unlikely to contain a radical change of topic. Since the candidate for discourse topic is again referred to, the gap is likely to be a long segment with a related topic and many subsegments. Even if it is not, a relatively short gap within the dialogue should be more conveniently analysed as a digression segment.

Absences of tokens at the beginning of a dialogue are more likely to contain a separate fragment. It is not unusual that dialogues start with a distinct topic which acts as a preamble to the actual subject dealt with in a dialogue. Absences in the end of the dialogue should be treated more sceptically, as the possibility of a related element is much higher, and it is not easy to predict what might or might not be related to a certain topic, because of the links specific to the dialogue situation. Dialogue-specific connections are much more unlikely to be confidently developed in the beginning, but, in a corpus such as the LLC, sampling may render this principle invalid. The analysis of Dialogue 1 will be carried further in order to illustrate fragment detection.

As mentioned before, the procedure for the identification of the discourse topic for a given dialogue should spot the presence of two distinct discourse topics within the same dialogue, although this may require careful probing. The initial forty-seven lines of Dialogue 1 present a relatively hard challenge, as the passage is short, making global effects difficult to spot. Going back from the first occurrence of *bibliography* in line 0047, it is easy to see that 0045-0046 are introductory lines where one of the participants announces his intention to actually get to the point of the conversation. It is also important to note that the first line in the dialogue which actually connects to what follows is 005, since the four lines preceding it are lost in terms of topic continuity, as a consequence of the sampling technique used in the London Lund Corpus. From 0044 to the beginning, the subject of conversation does not seem to bear any relation to what comes after 0044. In order to make sure, the approach used to the whole dialogue was repeated with this fragment, that is, a word frequency count was run to see what it could reveal.

Frequencies of suitable words for topical roles are very low in a short stretch like this. Once grammatical items and auxiliary verbs are discarded, the first candidate is *typing*, which, in the present circumstances, must be taken into consideration. It would be known that the item is a verb, as tagging is assumed. It occurs three times here and no more throughout the whole dialogue. As the threshold for words suitable for topical roles is lowered to two and then to one occurrence, the pattern is repeated, that is, they occur here and nowhere else, except for *work*, but that is a rather common word. Indications of a separate fragment with a radically distinct topic are now fairly strong. However, the size of the fragment would be quite small, which is unusual. It is also difficult to make sure because there are thirteen occurrences of *it* between the beginning of the dialogue and 0045, which means that any of the suitable words could be referred to several times by the pronoun. In a short stretch like this one, two or three references may make all the difference, as all suitable words vary between one to three tokens.

The next move would be then to check the distribution of the suitable items, but this is not likely to reveal much in a short passage, where some of these items appear only once. Nonetheless, *typing* appears three times close to the beginning only. Subjects are *I* and *you*, which are not suitable candidates. Objects are *it*, *final copy*, and *it* again, but the first object *it* occurs in the second line and there is only one possible antecedent in the first line. The item *thesis* emerges as a strong candidate in spite of the single token. It is the first suitable item to be introduced. The utterance is a question, which is a typical way to select a topic in dialogues, and appears at the very beginning. If selectional restrictions are considered, chances are that *thesis* is referred to at least three times through *typing*. There is also one token of *submitting* with a *it* for object, likely

to be another reference to *thesis*. The lexical item seems to be then a fairly acceptable working hypothesis, although it has been necessary to consider possible pronoun referents to help with the analysis, which should be avoided.

At the other end of Dialogue 1, there is a long absence of tokens for *bibliography* from 0891 (0910 if lemmatised tokens are considered) to the end. Could that mean a change in discourse topic? It does not seem likely. *Ford* occurs at the very end and it is a related element which was considered as a choice for discourse topic, appearing throughout. On the other hand, there is a 311-line gap between this occurrence of *Ford* and the preceding one, as pointed out before. Among the other tokens, there are the *University of London* references, linked to *Athlone Press* and the funding of *the bibliography*, and *University Microfilms*. Two of the tokens for *English* also appear within this stretch. There seem to be indications that the conversation is rambling about, which is not surprising after all the talking. There is an isolated reference to *Oxford Press*, and the token *Xerox* appears nine times between 0796 and 0977.

The final decision would have to be postponed. The analyst would have to read the dialogue before making it. The mild rambling hypothesis is confirmed on reading. The participants have already exhausted the matter and are adding other possibilities of funding to the conversation without much objectivity. Occasionally there is a reference to one of the previously mentioned agencies. Offset printing is discussed for a while as an alternative, with a short segment on the business practices of Rank Xerox. The fact is that there would not be any clear signs in the frequency count, and the distribution would reflect the rambling nature of the talk, preventing conclusions. A more detailed analysis in search of segment topics might yield results, but this is discussed in the next section.

Concerning the decision to split a dialogue into fragments, it should be weighed against considerations on size and position in the text. The procedure is shown below.

1. Analyse the patterns of lexical items in the dialogue, studying frequency, distribution, and first-appearance position of the chosen discourse topic in order to spot possible fragments.

2. Short fragments — less than 300 lines — should be avoided.

3. A digression segment is more appropriate than a separate fragment if a discourse topic resumes after a gap with an unrelated topic.

4. Shorter fragments are acceptable at the beginning of the dialogue, and possibly at the end, but less so.

The next step is then to consider the procedure for the segmentation of a dialogue according to topic continuity.

3.1.3 The identification of segment and subsegment topics

In the word frequency count for Dialogue 3, shown in 3.2, *doctor* appears as the second most frequent element. However, as explained above, the lexical item was discarded as a possible discourse topic due to its uneven distribution, and eventually *will* was chosen as the most appropriate solution. The item shows a more evenly distributed pattern of appearance, expressed in the virtual equivalence of the means. The other candidate for discourse topic, *mother*, also presented an even distribution and a much higher frequency rate (60 tokens), but it was nevertheless discarded on the basis of the preference for inanimate objects rather than persons.

As the decision was not clear-cut, the patterns of occurrence for *will* and *mother* were investigated in detail, showing that the longest gap for *will* begins just before the first appearance of *doctor*. The stretch where *doctor* tokens appear includes the 283 lines of the *will* gap plus 237 lines where the lexical items co-occur. This seems to suggest a long *doctor* segment, probably with several subsegments. It is also noticeable that there are twenty-five tokens of *phone* as a verb,

twenty-three of those in the 0160-0615 range. Therefore, chances are that an episode involving the *doctor* and a *phone* call is of some importance to the global topic *will*.

Of course there is no guarantee that things have in fact developed this way. As said before, these procedures cannot replace the analyst in regard to coherence matters. However, a careful exploration of the data on frequency rate, distribution and order of appearance may yield a great deal of information. It seems useful to map these exploratory moves into the procedure for segmentation. This may include local frequency counts, using 40-line chunks as a default. The size of the chunk is not meant to represent the probable size of a segment or subsegment. Subsegments are typically shorter and segments may be shorter or longer, with or without subsegments. The 40-line chunk is a snapshot of a given stretch of discourse chosen on the basis of observed global frequency and distribution patterns. The frequency count program has been customised to include not only a frequency count for the whole file but also frequency counts for every 40-line chunk in a file. The size of the chunk can be easily changed for further exploration by means of a simple command.

Let us suppose the analysis of Dialogue 3 had just been completed for the purposes of discourse topic identification. The working hypothesis for discourse topic is *will*, *mother* is a discourse thematic element, and *doctor* is also an important element which dominates a large stretch of the dialogue in some sort of relation with the verb *phone*. A few other elements present interesting effects, such as the lemmatised verb *read*, with twenty-five tokens, fourteen of them in the 0067-0115 range. There is a good chance that *reading of the will* will appear as a segment topic here. Two names, *Kay* and *Coleman*, show high frequencies, the latter being highly concentrated in a short stretch, 1103-1288, and so likely to be the topic of a segment developed within this passage.

Other effects might be mentioned, creating an interesting map of the topical organisation in the dialogue, but it would not be possible to be sure about topics and certainly quite impossible to make any statements about boundaries between segments and further divisions into subsegments. The frequencies for 40-line chunks will not achieve this either, but they will exert the same sort of sobering influence that the global frequencies did, forcing the analyst to aim at specific items occurring explicitly for topic selections. Looking at the first 40-line chunk in Dialogue 3, it is noticed that *lunch* appears three times in the initial chunk and *meal* shows twice. However, the first token is in 0012, which leaves still some room for a possible introduction for the discourse topic *will* at 0006 and *mother* at 0009. Reading the dialogue is the only way forward now. The tract of dialogue shown in (7) below is the 40-line chunk in question, with ten lines added.

(7)
0001 A Mr Potter
0002 A did you
0003 A arrive
0004 A about two o'clock
0005 A on the Sunday
0006 A the date the will was signed
0007 B yes
0008 A and did you go
0009 A and see your mother straight away
0010 B yes I did
0011 A what was she then doing
0012 B she was having her lunch
0013 A what about the brandy bottle
0014 A where was that
0015 B I don't know

0016 B I didn't s- I didn't see
0017 A you didn't see it
0018 B well
0019 B no I didn't
0020 B I I I all I know
0021 B was my mother was having her lunch
0022 B when I arrived
0023 A and
0024 A how did she seem then
0025 A at two o'clock
0026 B well
0027 B she seemed all right
0028 B I think she was a little tired
0029 A and how long did it take
0030 A for her to complete her lunch
0031 B oh I would think
0032 B probably
0033 B fifteen minutes
0034 A was it any a meal of any substance
0035 B she had erm chicken
0036 B she didn't eat very much of it
0037 A did you sit with her
0038 A whilst
0039 A she completed the meal
0040 B I was in the room
0041 B while she was having it
0042 B yes
0043 A and then uh did she have it on a tray
0044 B yes
0045 A somebody took the tray out presumably
0046 B er my wife took it out
0047 A and uh that's then about two fifteen
0048 B uh yes
0049 B i- yes
0050 B it would be

The dialogue is a bit more orderly than most in the sample, since it occurs in a courtroom situation. On the other hand, there is a high level of shared knowledge which permits new local topics to be introduced quite abruptly. This gives the beginning of a dialogue a choppy quality. The *lunch-meal* effect detected in the 40-line chunk frequency does not take hold as a topic right after the first appearance. It is introduced in 0012, but the *brandy bottle* is a new topic, signalling the start of a new segment. The *lunch* topic resumes in 0021, but again a question about *mother*'s

condition interrupts the development. The new resumption with the question in 0029 finally elaborates the topic *lunch*, with questions on duration, substance, the participant's presence during the meal, and the use of a tray, characterising the segment-subsegment organisation. The segmentation for the first 40-line chunk, extended to line 0050, is shown below with the topics for each unit:

0001-0007	'time of arrival'	segment 1
0008-0010	'seeing'	segment 2
0011-0012	'mother's lunch'	segment 3
0013-0019	'brandy bottle'	segment 4
0020-0022	'mother's lunch'	segment 3 (resumptive)
0023-0028	'mother's condition'	segment 5
	'mother's lunch'	segment 3 (resumptive)
0029-0033	'duration'	subsegment 1
0034-0036	'substance'	subsegment 2
0037-0042	'B's presence'	subsegment 3
0043-0050	'tray'	subsegment 4

Procedures for the identification of local and sublocal topics are likely to be less precise than the one specified for the identification of the global effects, as the effectiveness of figures such as frequency counts and distribution is strongly reduced by the small size of the stretches. As it is plain to see, some of the segment and subsegment topics involve adaptation of the actual verbatim forms of tokens in the dialogue. Nonetheless, the definition of such procedures appears to be useful for the standardisation intended.

The general map of topicality which the frequency and distribution data offer is not enough to support segmentation at the local level all the way, although it may offer important hints. Gaps without discourse-topic tokens in the beginning or at the end of dialogues may indicate a fragment with a distinct discourse topic. Items with a high frequency which concentrate exclusively on a certain stretch of the dialogue tend to be topics of large segments with many subsegments. Frequencies and distributions for specific stretches of various lengths may be requested to refine the map of lexical cohesion. As mentioned before, the research uses a customised version of the Berkeley HUM program for word frequency counting that also produces counts for dialogue chunks of any length, as specified by the analyst in a simple UNIX^{TM} command.

However, the actual segmentation, setting precise boundaries between segments and subsegments, has to be made manually. In order to guide the decisions, the adaptation of Hoey's method ([Hoe91]) mentioned above was combined to the analysis of exchange boundaries and coherence mechanisms — prospection and encapsulation — to be found in Sinclair [Sin93] and Francis and Huston [FH92]. These boundaries or initiating moves (see [Sin92]) typically select a new topic. The topic should then be analysed regarding the current segment topic. The analyst should decide whether it develops the segment topic (subsegment boundary) or is autonomously related to the current discourse topic (segment boundary). The procedure is described below.

1. Analyse the patterns of lexical items in the dialogue, studying frequency, distribution and concentration in ranges in order to spot potential topics for large segments

2. Analyse the frequencies and distributions in 40-line chunks and integrate results to information from the previous step

3. Request frequency counts for shorter chunks if necessary

4. Check manually by analysing the exchanges in terms of prospection and encapsulation mechanisms in order to spot boundary moves.

5. Analyse the boundary moves in order to establish whether they introduce:

- a segment: introduced topic is related to discourse topic but does not develop a current or previous segment topic, being a new local topic
- a subsegment: introduced topic develops current or previous segment topic, being clearly subsumed in this segment topic

6. Consider that the introduced topic may be best represented by an adapted form of a token in the dialogue

7. Prefer a new segment to a subsegment of doubtful subsumption

In a typical analytical situation, fragments — or at least clues of their existence — are very likely to be spotted during the process of frequency analysis. Thus, the analyst will have strong indications of where a fragment boundary might be, if any exist at all. It seems unlikely that a radical change in the discourse topic could go unnoticed after the frequency and distribution scrutiny. The difficult decisions will occur mostly in defining segment and subsegment boundaries. It may be unclear whether a new topic represents a break with the existing segment or is subsumed under an existing segment, even using the analysis based on the coherence mechanisms. The local frequencies guarantee that only a restricted universe of elements will be taken into consideration, but a degree of analyst agonising is inevitable.

Having in mind the rank scale of descriptive units for discourse analysis presented in Sinclair [SC92], it can be said that segments tend to coincide roughly with transactions, subsegments with exchanges, and fragments with interactions. However, this should not be understood strictly, because the authors claim to have developed a structural model for the analysis of discourse which does not rely on any semantic criteria. The segmentation model used here makes no such claim, as it is based on notions such as topic continuity, topic shift, global topic, local topic, and subordinate or sublocal topic. The procedures spelled out above are an attempt to overcome the inherent difficulty of dealing with the notion of topic. It seems certainly possible that segments may often be more easily identified with exchanges rather than transactions.

A second problem is the precise location of the boundary move. As noted in [Sin93], a sentence can relate to a previous one through the mechanism of encapsulation and, at the same time, relate to a subsequent one by means of prospection. Different elements in the sentence actually realise the coherence mechanisms. This may also happen in spoken discourse, which causes the analyst to be often faced with a decision as to where the boundary utterance should be placed, whether at the end of a unit or in the beginning of the new one. When an utterance "looks both ways" as described, it has been decided that it should be placed in the beginning of the new unit.

There are also discourse units — either segments or subsegments — which are clearly identifiable as such but do not have good candidates for the function of segment topic, due to specific features of discourse. This is particularly common when a new topic is introduced anecdotally, with the speaker describing a hypothetical conversation pretending to be the person involved, as a way of conveying meaning. The example below may help clarifying the sort of problem at issue.

(8) **B:** do clients ever say uh look Mr Chatwick let me give you um five hundred pounds or something

A: yes

B: and instead of you ringing me up all the time I will take this as as uh merely gambling money and you play it um and uh don't just have commission but let me give you ten per cent or something are you allowed to do this

A: yes yes we have a not quite under those terms well we would be allowed to do it but I don't think we would erm we have a thing

called a discretionary service um whereby people sign a little chit and that we deal for them without telling them

In the excerpt above, the introduction of a new topic is evident, as the participants are noticeably talking about something else before **B** begins describing the imaginary situation he wants to discuss. However, as the speaker himself is unsure of the way things actually happen, there is no concise way to define the topic of the segment before **A** answers. Even then, things are not made much easier for the analysis. Although the phrase *discretionary service* might be an appropriate solution, the most salient entity in the segment is not the service but the fact that the broker will be dealing for the clients without telling them. It is therefore acceptable to nominalise a chunk of speech — for instance, *dealing without telling the clients* — and use it as the segment topic, even when this requires a degree of adaptation. This option, however, should be used only when the more straightforward techniques fail.

3.2 The classification of anaphoric relations

There are several difficulties which are inherent to the classification of anaphoric relations. The annotation scheme was designed in such a way as to offer solutions for the problems of classification. The series of examples below is meant to describe these problems (all examples are taken from the LLC), thus concluding the discussion on methodological problems before the actual annotation scheme is described.

(9) **B**: erm in the sort of general outline that I sent you of the of the project how did it strike you

A: oh I think it's good

The antecedent for the first anaphor in example (9) can be straightforwardly identified using syntactic information. An algorithm such as Hobbs' [Hob86] would be able to handle the reference above. The second anaphor creates a chain of reference — a highly frequent phenomenon in dialogues (see [HH76]; [Bib92]). It seems reasonable to assume that Hobbs' 'naive' algorithm could be adapted to identify the antecedent for the second occurrence as well.

It is important to note, however, that finding an anaphoric personal pronoun which had its antecedent in the same sentence demanded some search through the corpus material. Example (9) cannot be said to be a typical case of intra-sentential anaphora as well. Regarding personal pronouns, the frequency rate of cross-sentence anaphora is higher than the one for within-sentence anaphora. In fact, it is difficult to identify the structural sentence — as usually understood — when the corpus data are made up of real-life dialogues. One form of anaphora which appears to be very common in dialogues is the reference within an adjacency pair (see [SSJ74]), such as the occurrence in the example below:

(10) **A**: how's the thesis going

B: uh I'm typing it up now, typing up the final copy

This sort of anaphoric reference can also be handled by an adaptation of Hobbs' 'naive' algorithm. However, one relatively harder problem to be dealt with is exemplified by the anaphoric nonpronominal noun phrase at the end of **B**'s utterance in example (10). The noun phrase refers to *thesis*, or, more precisely, its *final copy*, an implicit antecedent. Knowledge of academic work is needed to establish the link. This is usually called **world** or **experiential** knowledge. Recency or syntax will not be enough. It is therefore important that the annotation scheme provides means to record whether or not the antecedent has been previously introduced in the discourse. Example (11) below raises a different kind of problem regarding antecedents.

(11) **A**: you didn't know Mr Coleman was her solicitor until after she'd

signed the will

B: well it didn't really register that Mr Coleman was her solicitor to me

Example (11) highlights the fact that the antecedent for a given anaphor can be a discourse chunk instead of a single phrase. This is an occurrence of what is often named as the anticipatory *it* (see [QGSL85], section 2.59). The pronoun stands for a clause which is subsequently introduced. Hobbs explicitly states that his 'naive' algorithm cannot handle this kind of anaphora. Anaphors which may refer to chunks of discourse include demonstratives and do-so anaphoras, inter alia. The identification of the precise chunk being referred to can be quite challenging for an anaphora interpreter in an NLP system. The annotation scheme, therefore, should also include ways of codifying the occurrence of discourse-chunk antecedents.

There are occurrences where *it* appears without a referential function, or, at least, arguably so. These include collocations — see explanation below and definition in 4.4.13 — such as the one in example (12), and the so-called weather constructions. The fact that a typically anaphoric word, such as *it*, may be nonreferential must be addressed as well. Nonreferentiality should be made part of the codifying options in the annotation scheme. Other typically anaphoric words, such as *that*, can also be nonreferential.

(12) **B:** I hope you'll accept my word on this

A: yes

B: because I I mean it

A: all right I will

The distinction between a referential and a nonreferential pronoun is nontrivial. Many borderline cases are likely to appear as the analysis of corpus data progresses. As it often happens whenever corpus data are being analysed, the tokens collected are more easily defined as a continuum than as members of clear-cut categories. Thus, there are pronouns which are clearly referential, such as those in examples (9) and (10). However, the notion of anticipatory *it*, as applied to (11), is not such a consensus in all cases. In [QGSL85], the *it* as an anticipatory subject in cleft sentences is characterised as a pronoun that 'arguably has cataphoric reference' (section 6.17), although the authors seem inclined to acknowledge its referential value.

According to the same authors, occurrences such as the one in example (12) are 'the best case for a completely empty or "nonreferring" *it*' (see [QGSL85], section 6.17, note **a**). The actual interpretation of the utterance would be presumably *I am sincere in what I am saying*. No referent for *it*, either present in the discourse or inferred, is used in the interpretation. It seems safe then to classify example (12) as nonreferential. This approach suggests then that referentiality is to be measured according to the necessity of identifying a referent as a requirement for semantic interpretation. One intuitively reasonable way of testing whether this identification is required is to check for possible referents which would allow a more plausible interpretation of the utterance than the idiomatic resolution used in example (12).

If this line is adopted, one obvious way to perform the check is to replace the pronoun with the proposed antecedent in order to test whether the utterance is then satisfactorily understood. In example (11), such a substitution results in an acceptable utterance, although an arguably stilted and unlikely one in spoken language. In other cases, however, the antecedent, as identified in the discourse, may not produce such an acceptable outcome, as in example (13).

(13) **B:** it was very shortly after that interview that I sent my circular letter around to various scholars and I sent you a copy

The utterance derived by placing the that-clause in subject position is extremely awkward and very unlikely in real speech. Nonetheless, the structure is quite similar in many respects to exam-

ple (11), and the experience of annotating corpus dialogues shows that a variety of intermediate degrees of stiltedness and acceptability exist. There are also cases in which the acceptability of an utterance derived by the replacement of anaphor with its proposed antecedent is enhanced by a minor adjustment, as in (14) below, where changing *that* for *what* produces a much better result. This means of acceptability improvement is typical of cleft sentences, which have correspondent pseudo-cleft sentences in most cases.

(14) **B:** it's the academic structure of the university that that uh we're concerned about

It is true, however, that, once the idea of adjusting the proposed antecedent is made acceptable, it becomes necessary to determine the extent to which these adjustments can be said to be acceptable. A line must be drawn at a given point beyond which the proposed antecedent is to be considered as a creation of the analyst that cannot be correctly claimed to be available for reference. Moreover, there are cases in which an antecedent can be arguably identified as a vague generic noun phrase, like in example (15) below. These cases also appear in the dialogue with various degrees of vagueness.

(15) **B:** but then you see it's uh so strange I put my bibliography to the Oxford Press and I mean it's the most obvious press (...) and erm I don't know why Oxford turned it down

The first token of *it* can be said to be interpretable as referring to something like *the state of affairs* or *the situation*. This of course opens the possibility of analysing tokens of *it* in weather constructions as referring to *the weather*, and those occurring in expressions denoting time as referring to *the time*. As these decisions were part of the daily routine of annotation, it was necessary to define a standard for the attribution of referentiality. This standard, similarly to the procedures for the identification of topical roles, is not intended to eliminate controversy. It provides, nevertheless, a definite way of analysing tokens of typically anaphoric words and decide on their referentiality status. The standard is spelled out in section 4.2 in Chapter 4, where the actual categories used to classify anaphora cases in the study are described.

Whichever standard is chosen, it is important to establish whether certain collocations of typically anaphoric words — such as *it* and *that* — are regularly interpreted in a way which is different from the expected, assuming that the expected interpretation is an anaphoric reference to an antecedent identified by means of an algorithm such as Hobbs or some equivalent processing strategy used by humans. One could then imagine 'a mint of phrases', to use Kjellmer's expression (see [Kje91]), for the purpose of anaphora processing. The *it* in cleft sentences is very likely to be interpreted as a cataphoric reference. The object pronoun in collocations with *mean* may be often nonreferential. If this is true, it will be certainly very helpful to record these phrases with their typical resolution as a way to avoid frustrated attempts to handle them in the expected way.

The analysis of corpus material has also shown that, once discourse elements are well established, distant anaphoric reference is not uncommon and certainly possible. Referential chains may also be interrupted without clear linguistic clues, provided certain associations are stable enough to ensure correct identification of the antecedent. Thus, as in example (16), pronouns which are apparently linked in a chain may have different antecedents. These associations can only be established by full discourse processing. Example (16) gives an idea of how anaphoric reference may depend on discourse information for resolution.

(16) **A:** and when you were reading the nineteen sixty-four will did mother at the same time have the nineteen sixty-one will with her

B: it was on her bed but it wasn't open in front of her

A: when you read it did you read the whole thing through

The antecedent for the first and second occurrences of *it* is *the nineteen sixty-one will*, but the third occurrence, in A's utterance, refers to *the ninety sixty-four will*. At this point in the dialogue, the fact that one will has been read by **B** for *mother* to hear and the other has not is firmly established. There is no risk of the reference being misunderstood, in spite of recency and usual chaining processes pointing to a different resolution. The effect of topicality on anaphoric phenomena seems to be the best hope for an effective handling of these occurrences. In a nutshell, the hypothesis would be that those highly salient elements are the ones which can be referred to in a way that violates the usual constraints on recency and chaining.

It should be noted that all occurrences of anaphora in examples (9) to (16) involve *it* as an anaphor. Information needed for the processing towards resolution is nonetheless distinct. It is therefore not enough to classify anaphors according to traditional part-of-speech categories, if the purpose is to understand the processing involved. Different tokens of the same personal pronoun or any other kind of anaphor may be resolved with the use of distinct processing strategies. At the same time, a classification based on part-of-speech categories is equally needed, because the anaphor is the element in an anaphoric relation which triggers the processing required for the identification of an antecedent. The annotation scheme should thus include a classification of processing strategies, as well as one for the types of anaphor, so as to characterise the differences in processing which may exist between two tokens of anaphora even when the anaphoric terms are identical or of the same type.

The elements which should be codified into the annotation scheme, to sum up, include: the topical roles described above, for the purposes of both segmentation and identification of the antecedent status for each occurrence of anaphora; a classification of the anaphoric term which would allow clues in the verbatim form of discourse to be recognised as a signal of an anaphoric reference; a classification of the antecedent as implicit or explicit, which should also include the nonreferential option; a definition of the processing strategy involved in the resolution of each occurrence, so that tokens of the same anaphor are not simply lumped together under a label, such as **personal pronoun**. The next subsection describes the general features of the annotation, which represents an attempt to deal with the various problems discussed above.

3.3 Features of the annotation

Edwards [Edw92] proposes a set of six principles to ensure readability in an annotated transcript. This section discusses these principles briefly before presenting the annotation itself, in an attempt to show the motivations underlying the choices of form made. These choices also bear in mind the characteristics of the anaphoric world which were defined as requirements for a successful annotation scheme in the previous section.

The first principle is called **proximity of related events**, meaning that types of information which are more closely related to each other should also be spatially closer. As explained before, the annotation aims at codifying information regarding both the discourse level — involving topicality and segmentation — and the anaphora level, where each case of anaphora is classified. Consequently, one can interpret the requirements of the principle for the discourse units — fragment, segment and subsegment — as being satisfied by annotation entries at the beginning of each unit. Regarding each anaphora token, one could think of solutions which would place the classification immediately before or immediately after the anaphor which triggers the resolution process.

The second principle is the **visual separability of unlike events**. Edwards intends **unlike events** to mean 'qualitatively different types' of information, for instance, spoken words as compared to researcher comments or entered code. However, the annotation scheme planned involves qualitatively different information within itself. It seems useful to separate discourse-segmentation code from anaphora-case code. Information related to fragments, segments, and subsegments can

be placed in separate lines at the beginning of their respective units with an identical nonalphabetic character at the beginning of each line, so that the segmentation information can be unequivocally recognised. Code for each anaphora case can be entered immediately after the anaphor in a transcript.

The principle of **time-space iconicity** concerns the ordering of events in the transcript, which was previously and efficiently settled in the LLC. The transcription of the Portuguese data follows the same principles whenever possible. The placement of the information about the discourse unit before the unit seems to be more consistent with the **logical priority** principle, which requires that logically prerequisite information be placed before the utterances concerned. The unit is visualised more easily with the line stating the type of unit — fragment, segment, or subsegment — the number, in a sequential ordering, and the topic at the beginning of the discourse unit. References to the topical role in the classification of the antecedent would be impossible to interpret without awareness of the current topics at the global and local levels. Thus, the analysis codified in the annotation is more easily understood by a reader if the discourse-segmentation level of information is known from the start.

The annotation scheme tries hard to respect the principle of **iconic and mnemonic marking**, which requires that the code used be related to what they stand for in a recognisable way. However, the categories needed to classify the types of anaphor and the processing strategies involve unusual concepts which are likely to require the use of a key for perfect understanding. Thus, the abbreviations used are as mnemonic as possible, but it takes some acquaintance with the annotation to recognise the code without resorting to a key. Improvements would be likely to jeopardise **efficiency and compactness**, as too many items of code would have to be entered in order to make the entries truly mnemonic. The trade-off resulted in some symbols having six letters, for instance, but never more than that. Moreover, a large majority of the symbols used require two or three letters only.

In order to cover all the information requirements previously defined, the annotation for each case of anaphora includes four properties. The first one is the type of anaphor, which classifies the anaphoric term according to categories which coincide to a significant extent with traditional parts of speech. The full set of categories is listed in 4.1. The second property is the type of antecedent, which concentrates on the implicit-explicit dichotomy with a few additions, such as the nonreferential option, which are defined in 4.2. The third property defines the topical role of the antecedent according to the topicality hierarchy described above. The distinction between single-phrase and discourse-chunk antecedents is also codified here. The fourth property is the processing strategy, which is classified according to categories listed in 4.4.

The annotation scheme enters the four items of code for each one of the four properties sequentially between brackets immediately after the anaphor. Each item of code is separated from the other by a semicolon. As the anaphoric references are far more common than the cataphoric ones, only the latter are signalled in embedded brackets immediately after the code for the type of anaphor, the first one to be entered, but before the semicolon which separates it from the code for type of antecedent. An example is shown in 3.3.2.

3.3.1 Topic in the annotation scheme

All information at the discourse level is entered in lines marked with a single asterisk in the first column. Information about the discourse topic is entered before the dialogue or dialogue fragment for which it is the topic. If there is only one discourse topic for the full dialogue, the expression SINGLE FRAGMENT is annotated. Otherwise, the ordinal ranking of the fragment is specified (first, second, etc.). An example is given below.

* (FIRST/SECOND/SINGLE) FRAGMENT — 'bibliography'

Information about segments and subsegments is also entered before they begin, thus marking the boundary. Lines containing segment and subsegment information are also marked with a single

asterisk in the first column. This is followed by the letter **s** for segments or **ss** for subsegments, together with a number which identifies the unit sequentially by order of appearance. A subsegment mark also specifies which segment it is part of. The marks are followed by a phrase which specifies the topic for that unit, as identified by the procedures in 3.1. The annotation for segment and subsegment boundaries is shown below:[3]

(17)
* s19 'B's statement through solicitors'
A: your solicitors furnished a statement made by you to the defendant's, is that within your knowledge
B: yes yes
A: you know that
B: yes
A: mm
* ss1/s19 'contents of statement'
A: did you know that in that statement furnished by your solicitor to the defendant's, it's stated: both Elsie and I had suggested this to mother before I phoned her doctor who was out but arranged with the receptionist that I'd phone him early next morning from my home This I did
B: that is so
* ss2/s19 'day of B's phone call'
B: but that was the Thursday - er before

Segment and subsegment topics may resume in a dialogue which revolves around the same discourse topic. Whenever this happens, an **r** is placed before the segment or subsegment mark. Thus, a resumptive segment will be marked **rs3**, and a resumptive subsegment **rss3**. Suppose segment 19 has a segment topic which is developed into subsegments as a sequence of five different subsegment topics. On the sixth subsegment, subsegment topic three resumes. The discourse-unit mark next to the asterisk will be **rss3/s19**. Other possibilities of topic development occur. After our hypothetical **rss3/s19**, the topic of previous segment 15 may reappear with a direct reference to the segment topic and not to a subordinate topic. This would be marked **rs15**. It may be developed into a subsegment with a subsegment topic which had not appeared previously. This would be marked **ss1/rs15**. One of the previous subsegment topics may also resurface with the segment. Suppose the former first subsegment resumes. This would be marked **rss1/rs15**. The dialogue sample is thus fully segmented and annotated according to this scheme.

3.3.2 Anaphora cases in the annotation scheme

Each case of anaphora is annotated by inserting four slots of code between round brackets next to the anaphor token being analysed. The first slot contains the code for the type of anaphor. The second slot defines the type of antecedent. The classification for the third property, the topical role of the antecedent, is entered subsequently, followed by the category which specifies the processing strategy. Each one of these slots is separated from the other by a semicolon. An example is given below.

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

[3]The one-tone-unit-per-line arrangement and other features of the RLLC are edited here for reasons of space. Some punctuation marks are added as well, for the sake of easy comprehension.

* ss4/s38 'Captain's personal witnessing'

A: w- did he (SP; ex_222; thel; FtC;) say that he (SP; ex_222; thel; FtCCh;) had personally witnessed one (One_an; ex_1; dt; SetMb;)

B: well I could have been I could have been wrong there (AdvP; ex_116; p_sst; CK;)

In example (18), the first bracketed group of annotation defines the type of anaphor as **FNP**, which stands for nonpronominal anaphoric noun phrase. This is followed by **ex_222**, which classifies the antecedent as explicit and identifies it as number 222 in the referent list for the dialogue. The code **dthel** means **discourse thematic element**, a topical role attributed to entities of global saliency in a dialogue which are not the discourse topic. The processing strategy is specified as **LR**, code for **lexical repetition**, which means that a search backwards for a similar token finds one that is precisely the same as the anaphor, thus resolving it. The second bracketed group of annotation identifies the anaphor as **SP**, which stands for subject pronoun, with an explicit antecedent labelled as number 224. The antecedent is also a discourse thematic element (**dthel**) which is correctly identified by simply selecting the first candidate in a search backwards. The code for the processing strategy is therefore **FtC**, indicating that the successful strategy is a first-candidate search.

The next chapter describes the categories used to classify the anaphora cases according to each one of the properties claimed to be required for an adequate analytical annotation aimed at anaphoric phenomena. A compact listing of the categories with the corresponding codes for quick reference is given in Appendix B.

Chapter 4

Description of the annotation scheme

Once the topic roles have been assigned throughout the full extension of the dialogue, the analysis of anaphoric relations can start. A fine-grained classification, with a relatively large set of categories for each property, is used to characterise each token. Two of the properties — namely, type of anaphor and processing strategy — require a particularly detailed classification. The symbols used in the annotation are placed next to the category name. Each category is presented with one or more examples extracted from the corpora, except for some of those categories classifying the antecedent according to topical role, which are better described by the identification procedures in Chapter 3. Yet details concerning this property will be added, and examples will be given when needed. Examples may of course contain other anaphors that not the one being used as an example, but, for the sake of clarity, only the annotation for the anaphor in question is shown.

4.1 The type of anaphor

This property refers to the word or phrase which triggers the anaphoric link, that is, the visible item which requires the retrieval of another element in the text for its interpretation. Concepts such as zero pronouns or empty categories are not used for the annotation. This means that a verb without a phonetically realised subject is annotated as an anaphoric verb. In addition, a response form, such as *yes*, which requires the retrieval of a previous sentence for its interpretation, is annotated as a **reaction signal**, according to the classification of adverbs in [QGSL85], section 7.54.

In short, the annotation for type of anaphor records what is phonetically realised as conventionally as possible, having [QGSL85] as reference for English and [CC85] for Portuguese. Features related to the antecedent and the processing are not marked in the code for type of anaphor. This approach was chosen in order to simplify the mapping from a POS-tagged dialogue. It is nonetheless true that some categories used to classify the type of anaphor have been created for the purposes of this research. The code for the type of anaphor is placed in the first slot inside the brackets next to annotated tokens.

4.1.1 Nonpronominal noun phrase (FNP)

It includes all lexical repetition, repetition with modifiers added, part-whole and other semantic links, plus a variety of connections such as the implicit binding of *doctor* to *receptionist* in example (17) in 3.3.1. Another example is shown below.

(19) **B:** I don't know whether you have talked with Hilary about the diary situation

A: well she has been explaining to me rather in rather more general terms erm what you are sort of doing and

B: what it was all about yes

A: I gather you've been at it for nine years

B: erm by golly that's true yes yes it's not a long time of course in the uh in this sort of work (FNP; im_5; thel; SK;) you know

4.1.2 Anaphoric adjective (AdjAn)

The category classifies adjectives which require the retrieval of a clausal or noun antecedent for semantic interpretation. The anaphor may appear either in the comparative or superlative form (example (21)), as well as in the standard form of the adjective (example (20)).

(20)

A: was there any time between your arrival at two o'clock and your departure after she had signed the will when she had any alcoholic drink

B: no

A: are you sure (AdjAn; ex_162; p_st; VMm;)

B: I'm absolutely positive (AdjAn; ex_162; p_st; VMm;)

A: very good

(21)

A: I just took it out of the shelf that particular volume because it was the smallest book

B: mm mm

A: you know I just go into uh a stationer and buy whatever happens to be there you see and that happened to be the smallest (AdjAn; ex_29; dt; AM;)

There is a contrast between tokens of this type of anaphor in English and Portuguese. As there is no equivalent for the anaphoric *one* in Portuguese, the occurrence of adjectives as noun phrase heads, referring back to a noun, is the most frequent form of anaphoric adjective, whereas this construction is only possible in special situations in English. One example is shown below.

(22)

A: o peso menor que a senhora pode atingir

gl: the-MASC weight smaller that the lady can reach

tr: the lowest weight you should reach

A: são quarenta e seis e oitocentos

gl: are forty and six and eight humdred

tr: is forty-six eight hundred

A: quase quarenta e sete quilos

gl: almost forty and seven kilos

tr: almost forty-seven kilos

B: o menor (AdjAn; ex_147; sst; VMm;) né ?

gl: the smallest, not is ?

tr: the lowest, isn't it ?

4.1.3 Subject pronoun (SP)

All occurrences of *it*, *he*, *she* and *they* are annotated, including those which may be considered as not truly anaphoric, such as *it* in weather constructions and collocations. Also all occurrences of *ele*, *ela*, *eles* and *elas* as subject pronouns in Portuguese. Some occurrences of first and second person pronouns are annotated when they appear in verbatim reproduction of speech, referring to third parties that not the dialogue participants. The example below shows two occurrences of annotated subject pronouns. The second one is a case of first-person anaphoric pronoun in verbatim reproduction of speech.

(23) **A:** erm - uh they (SP; ex_149; dthel; FtC;) suddenly say oh I've (SP; ex_149; dthel; ScRf;) got another two thousand

4.1.4 Object pronoun (OP)

As above, for the object counterparts. Object pronouns occurring as part of a contraction with a preposition in Portuguese, as in example (25) below, are included in this category.

(24) **B:** I don't know whether you have talked with Hilary about the diary situation

A: well she has been explaining to me rather in rather more general terms erm what you are sort of doing and

B: what it was all about yes

A: I gather you've been at it (OP; ex_267; dthel; FtCCh;) for nine years

(25) **B:** eu devia até ter trazido uns exames

gl: I should even have brought a-MASCp exams

tr: I should really have brought some exams

B: lá que eu tenho, não é, mas

gl: there that I have, not is, but

tr: that I have, isn't it, but

A: a senhora lembra de cabeça

gl: the lady remembers of head

tr: do you know by heart

A: algum deles (OP; ex_122; st; FtC;)

gl: some of-they-CONTR-MASCp

tr: any of them

4.1.5 Demonstrative (De)

All occurrences of *this*, *that*, *these* and *those* as pronouns in English, and of *isso*, *aquilo*, *este*, *esse*, *aquele* in all inflections in Portuguese. References to discourse chunks and tokens in collocations are annotated. In the example given below, *that* refers to the clause *you've been at it for nine years*.

(26) **A:** I gather you've been at it for nine years

B: erm by golly that's (De; ex_3; p_st; CK;) true yes yes

4.1.6 Determinative possessive (Pos)

All occurrences of third-person possessives — and those occurrences of the other possessives which appear in verbatim reproductions of speech as specified in 4.1.3 — as determiners.

(27) **A:** the the the the sort of Harold Macmillan the the um Harold Nicholson type who write their (Pos; ex_22; sst; FtC;) diary because they are aware of having their (Pos; ex_22; sst; FtC;) pulse on the on the goings on of the time

4.1.7 Independent possessive (PPos)

All occurrences of third-person possessives — and those occurrences of the other possessives which appear in verbatim reproductions of speech as specified in 4.1.3 — as noun phrases. Tokens of third-person pronouns are typically annotated as two cases of anaphora signalled by one single word. The first reference is to the possessor and the second one to the omitted possessed element. The tokens of possessive pronouns for other persons are annotated as one occurrence only, as referring to the omitted possessed entity.

(28) **B:** oh there is the analog thing yes

A: the Middle English analog er one of Peter Harringay

B: kept in Cirencester yes

A: no er Freeman's now all right he's gone ahead and Nelson in fact are doing his (PPos; ex_58; thel; FtCCh;) (PPos; ex_57; st; VMm;) aren't they

4.1.8 Adverb of place (AdvP)

All occurrences of adverbs of place which refer anaphorically.

(29) **B:** I'm sure my lord my mother didn't phone the doctor at three o'clock on that day

C: or at all while you were there (AdvP; im_13; thel; DK;) that afternoon

B: or at all while I was there (AdvP; im_13; thel; FtCCh;) that afternoon

4.1.9 Adverb of manner

Tokens of this type of anaphor were only found in the Portuguese sample. All cases are tokens of the word *assim*, which can be translated as *like that* or *so*, although the word is not used as an anaphor in the Portuguese equivalents of the types of anaphor described in 4.1.19 and 4.1.20 below.

(30) **A:** aí é aquele negócio que eu falei para a senhora

gl: there-NS is-EX that-MASCs business that I spoke-1sts to the lady

tr: That's what I was telling you about

A: o cálcio não vai ser tão bem absorvido

tr: the-MASCs calcium not go-PRES3rds be-EX-INF so well absorbed-MASCs

gl: calcium is not going to be so well absorbed

A: se a senhora fizer as refeições assim (AdvM; ex_66; dthel; DK;)

gl: if the-FEMs lady do-FSU the-FEMp meals so

tr: if you have your meals like that

One possible equivalent in English for many but not all usages of *assim* as an anaphor is the the pro-form for process adjuncts *thus* (see [QGSL85], section 8.78, note **b**). However, since it is largely a formal word in English, no tokens were found in the sample of English dialogues.

4.1.10 Adverb of time

This type of anaphor also occurred only in the Portuguese sample, as adverbs of time are not normally used as responses in English.

(31) **A:** e a senhora ainda sente a ... ainda sente

gl: and the-FEMs lady still feels the-FEMs ... still feels

tr: and you still feel ... still feel

A: aquela vontade de ir no banheiro durante o dia muitas vezes ?

gl: that-FEMs volition of go-INF in-the-CONTR bathroom during the-MASCs day many times ?

tr: the need to go to the toilet many times a day ?

B: Ainda (AdvT; ex_9; p_st; VMm;)

gl: Still

tr: Yes

Adverbs of frequency (AdvF), such as *nunca (never)*, and **adverbs of exclusion (AdvE)** — typically *só (only)* — can also be used anaphorically in responses in Portuguese.

4.1.11 One-anaphora (One_an)

All occurrences of *one* as an anaphor used as a substitute for a count noun, as distinguished from the use of *one* as an anaphoric numeral in a noun phrase where the head is omitted rather than substituted. See 4.1.12 for an example of the second type and also [QGSL85], 6.54–55, for a characterisation of the distinction. There is no direct equivalent to the anaphoric *one* in Portuguese.

(32) **A:** did you explain to her even if you didn't show the nineteen sixty-one will the basic difference namely that do you realize mother you're giving me another twelve hundred and fifty pounds over and above what Maureen was getting under the other one (One_an; ex_128; dt; SetMb;)

B: no I said nothing like that to my mother

(33) **B:** uh so that anybody coming across the only record that is public namely the one (One_an; ex_34; thel; SetMb;) that's in the in the filing cabinet

4.1.12 Numeral (NUM)

All occurrences of numerals when functioning as pronouns, both in English and Portuguese.

(34) **B:** the the meaning of that little diagram is that everybody's got to do the central three (NUM; im_105; thel; Dx;)

4.1.13 Indefinite pronoun (IP)

All occurrences of indefinite pronouns which refer anaphorically. These correspond to the *of*-pronoun category of indefinite pronouns and exclude the compound indefinite pronouns (see [QGSL85] 6.45–48, for a definition of the distinction).

(35) **B:** er in order to complete it I will have to visit the major resources in the United States and uh several (IP; ex_11; thel; VMm;) in Europe

4.1.14 Wh-word (WHT)

These occurrences are related to the use of *wh*-words both as interrogative pronouns and as subordinators to introduce nominal clauses. Either the question predicate or the nominal clause introduced has to be retrieved from the preceding discourse, and the anaphor which actualises the reference is a wh-word.

(36) **A:** erm have you tried the Oxford Press which is an obvious one

B: er no we didn't erm and I

A: it seems such an obvious choice that I can't understand why (WHT; ex_53; p_sst; VMm;)

4.1.15 Prepositional phrase (PP)

Prepositional phrases can refer to clauses in the preceding discourse to which they are attached as adverbials, as in the example below, where the prepositional phrase is the complement of a copula. The annotation is entered immediately after the preposition which is the head of the prepositional phrase.

(37) **C:** were you with your mother the whole of that afternoon

B: yes I was my lord

C: uh not out (PP; ex_242; p_dthel; VMm;) of the room at all

B: I don't think I left the room at all

This kind of anaphora also occurs with prepositional phrases attached to other types of verbs as adverbials. In the example below, an interesting process of gradual build-up can be observed, where further information is requested by means of added clause elements which refer to the initial question without repeating it. First, the direct object is referred to by lexical repetition, instead of a pronoun, with a quantity partitive added, and interrogative intonation, which plays a role throughout. Then, the subsequent prepositional phrase functions as a request for further information on the manner of performing the action defined by the verb. Then a periphrastic interrogative pronoun combined with the noun phrase *turns* is added as a contrasting question, all without repeating the subject and the verb.

(38) **A:** did you read it aloud to her

B: I did yes

A: the whole will

B: yes

A: in (PP; ex_19; p_st; VMm;) one go or how many turns

Anaphoric prepositional phrases are more frequent and diverse in Portuguese, but most cases fit the response pattern exemplified for adverbs. One example is shown below.

(39) **A:** mas a senhora continua com a mesma com o mesmo sintoma ?
gl: but the lady continues with the same with the same symptom ?
tr: but do you still have the same symptom ?

B: Com (PP; ex_13; p_dthel; AM;) o mesmo problema
gl: with the same problem

4.1.16 Reaction signal (AdvR)

Those are anaphors which are characterised by a type of adverb — mostly *yes* or *no* — which is a response to a former utterance, often a question. They require the retrieval of the previous utterance for semantic interpretation.

(40) **A:** but you have applied er for monies I keep hearing wherever I go
B: yes (AdvR; ex_266; p_st; VMm;) and with no result

4.1.17 Operator (OPT)

This kind of anaphor is marked by a subject followed by an auxiliary verb. The verb phrase has to be retrieved from the previous discourse to allow interpretation of the subject-auxiliary sequence. The short answer is the typical — but not the only — form of the anaphor.

(41) **B:** well David Tate had a boy at Charterhouse
A: yes he did (OPT; ex_275; p_sst; VMm;)

(42) **A:** if we decided to make we've decided to take somebody into account say your mother and it is unlikely actually that we shall (OPT; ex_67; p_st; VMm;) Mr Chatlick

4.1.18 Anaphoric Verb (VerbAn)

This classification applies whenever components in the argument structure of the verb have to be retrieved for semantic interpretation. References to a subjectless infinitive clause which functions as a direct object are included in this category.

(43) **A:** British Academy I'm I'm sure for the first year or so will be taken up by people that they will have been wanting to help for years
B: yes wanting to help for years yes
A: er and indeed have been helping (VerbAn; ex_108; st; VMm;) in a very small way

The phenomenon is much more common in Portuguese, a language in which it plays a central role in the referring system. Two tokens of Portuguese anaphoric verbs are shown in example (44) below. In the first one, the direct object of the transitive verb *pedir* in its third-person past tense

form *pediu* is omitted. In the second one, both the subject and the object of the same verb form are omitted, so that the anaphor token is annotated as two cases of anaphora.

(44) **A:** a senhora sabe se tem algum exame de sangue da senhora
gl: the lady knows if has any exam of blood of-the-CONTR lady
tr: do you know if there is any blood test of yours ready

B: de colesterol , de (2syl) glicídio ?
gl: of cholesterol , of glucosides ?
tr: like cholesterol or glucosides ?

B: 'tava ... foi a foi a a doutora pediu (VerbAn; ex_132; st; FtC;) né ?
gl: was ... was the was the the doctor asked-PAST-3s not-is-CONTR?
tr: the doctor has asked for them

A: pediu (VerbAn; ex_120; thel; FtC;) (VerbAn; ex_132; st; FtCCh;) ?
gl: asked-PAST-3s
tr: she has ?

The same holds for catenative verbs and semi-auxiliary verbs, as in example (45) (see [QGSL85], 3.47–49, for definitions), which are not annotated as operators.

(45) **B:** you read Sir Gawain and the Green Knight uh or Chaucer period
A: no er I've read Chaucer yes but it means very little
B: mm did you have to (VerbAn; ex_114; p_sst; PI;) as part of the Tripos,

4.1.19 *So* anaphora (SoAn)

This category contains all occurrences of *so* which refer to a previously introduced clause acting as the object of the preceding verb. There is no direct equivalent of this type of anaphor in Portuguese.

(46) **A:** did you know the doctor had spoken to Mr Spackman the day after the funeral
B: uh no at least I don't recall having known so (SoAn; ex_83; p_st; VMm;)

4.1.20 Do-phrase anaphora (DPA)

All occurrences of *do*, usually but not necessarily followed by *so*, *this/that* or *it*, which refer anaphorically to a previously introduced verb phrase. Differently from operator anaphors, *do* is not an auxiliary here, but an actual pro-form which replaces the former verb phrase (see [QGSL85], sections 12.21-26, for a detailed account of the distinction). The example below helps clarify matters.

(47) **A:** I say no more if you want to recall the doctor you may do so (DPA; ex_41; p_sst; CK;)

If *do so* were not part of the utterance, the reference would be still perfectly understandable,

and the antecedent would be the same. However, the occurrence would be classified as an operator anaphor (see [QGSL85], section 5.1.). Examples below are also classified as do-phrase anaphors.

(48) **A:** I wasn't asked is the answer if it was a legal document and never mentioned having witnessed a will on an envelope had I done so (DPA; ex_115; p_sst; CK;) it would have been a lie

(49) **B:** this wouldn't work properly the first time and mother signed again going off the edge of the page mother then said I'm going to do it (DPA; ex_88; p_dt; CK;) again as I don't want there to be any trouble

(50) **A:** basically er you you must try and get into the equities I think and you must try and do that (DPA; ex_134; p_st; CK;)

In some occurrences, lexical *do* appears as an intransitive substitute verb, without combining with any of the pro-forms above (see [QGSL85], section 12.22). Such occurrences are classified as do-phrase anaphors as well, regardless of the fact that there is no actual phrase to speak of. In the example (51) below, *didn't* is an operator, and the annotation for the type of anaphor is **OPT**. On the other hand, *done* cannot be an operator. It is classified as a do-phrase anaphor (DPA), although there is no combination with one of the pro-forms.

(51) **A:** well you knew then that it was a will prepared by Mr Coleman the solicitor at Hove

B: well I didn't (OPT; ex_144; p_sst; CK;) uh yes I I would have done (DPA; ex_144; p_sst; CK;) if I'd have read that uh that it w- that Mr Coleman's name was on it

In their discussion of *do* as a substitute verb, [QGSL85] (sections 12.21 to 12.26) refer to the subtle distinctions in usage between *do so* and the other two forms *do it* and *do that*. Whereas the latter are clearly combinations of the transitive main verb *do* with the pronoun *it* or the demonstrative *that*, the *do so* construction is grammatically dubious, in the sense that the status of *so* as a pronoun or an adverb is arguable. These distinctions do not seem to be relevant for the classification created for the purposes of this research. The even more subtle distinctions in usage between *do it* and *do that* also need not concern the present discussion.

Some verb classes (as classified in [QG73], section 3.35) do not admit the combinations typical of do-phrase anaphors. The class called 'verbs of inert perception and cognition' in [QG73], to which *know* in the example above belongs, is one of them. The anaphoric use of forms of the verb *fazer* in Portuguese were also classified as belonging in this category, as it seems safe to consider them as equivalents. One example is given below.

(52) **A:** quando tiver queijo,

gl: when have-FSU cheese

tr: when there is cheese

A: a senhora come ou o queijo ou toma o leite

gl: the-FEM lady eats or the-MASC cheese or takes the-MASC milk

tr: you either eat cheese or drink milk

B: ah tá

gl: ah is-ST

tr: all right

A: nunca na mesma refeição

gl: never in-the-CONTR same-FEM meal

tr: never in the same meal

A: ou deixa para fazer (DPA; im_49; p_sst; VMm;) na próxima refeição

gl: or leave to do in-the-CONTR next-FEM meal

tr: or else do that in the next meal

4.1.21 Anaphoric non-finite clause (NFClAn)

This type of anaphor is quite rare. The antecedent is a noun phrase which typically contains a quantifier — often a numeral — while the subsequent non-finite clauses specify the objects introduced by the noun phrase. The annotation is entered immediately after the non-finite verb.

(53) **B:** in the hope that they would do two things firstly - to give (NFClAn; ex_68; p_sst; SetMb;) me uh ay- small Ford Ford Foundation travelling grant to visit a number of key centres and universities to explore the land so to speak - and uh when that has been done to submit (NFClAn; ex_68; p_sst; SetMb;) to them a full documented report with the backing of virtually every major library and every major philologist in the world to get them to give me a substantial sum of money to enable me to finish it

4.1.22 Anaphoric that-clause (TClAn)

This is also a particularly infrequent type of anaphor. It consists of a sequence of subordinate clauses in which the main clause — or a fragment of it in the case of embedding — only appears once before the first one in the sequence. The annotation is entered immediately after the subordinating conjunction or relative pronoun.

(54) **B:** but the private diary that you write or that (TClAn; ex_2; dt; Pl;) I write and which (TClAn; ex_2; dt; Pl;) we hope will always remain under lock and key

4.1.23 Linking verb (LV)

This form of anaphor involves reference — by means of an uncomplemented copula — to a predicative adjunct, which has to be retrieved for the interpretation of the copular sentence.

(55) **B:** these are semi-personal however the solicitor knows you by first name

A: yes well of course the bank manager's letters are (LV; ex_52; sst; VMm;)as well

As in all anaphors involving verb forms, this type of anaphor is much more common in Portuguese. Forms of the linking verbs *ser* and *estar* are used as reaction signals and tag questions both in contexts where these verbs appear in the preceding statement and in contexts where they do not. These verb forms, therefore, perform pragmatic functions in which their lexical meaning is partially or totally lost. The complex interactions between anaphoric reference and discourse

markers in these contexts will be the subject of a future paper. Two examples are given below. In the first one, the verb form *é* responds as expected to a question where the verb form is the same. In the second one, however, the verb form acts as an affirmative reaction signal which bears no lexical relation to the verb in the previous sentence.

(56) **A:** diminuiu novecentos gramas
gl: diminished-PAST3rds nine hundred grams
tr: you lost nine hundred grams

A: mas é uma boa coisa, né, em um mês
gl: but is a-FEMs good-FEM thing, not is, in a-MASC month
tr: but it is a good thing, isn't it, in a month

B: é (LV; ex_6; p_dt; FtCCh;)
gl: is
tr: yes, it is

(57) **A:** e aí você fez uma uma pequena cirurgia (2syl)
gl: and then you made a a small surgery
tr: did you have a small surgery then ?

B: é
gl: is
tr: yes, I did

4.1.24 Copula-plus-noun phrase anaphor (CopFNP)

This type of anaphor may refer to a specific copular subject left out in coordinate sentences, or to broad chunks of discourse, which may be more or less defined. Only one case — of the first kind — was found in the English-language sample.

(58) **A:** I just took it out of the shelf, that particular volume because it was the smallest book
B: mm mm
A: you know I just go into uh a stationer and buy whatever happens to be there you see and that happened to be the smallest and was (CopFNP; ex_29; dt; Pl;) the most convenient to carry

This type of anaphor is far more common in Portuguese, with many occurrences of the second kind, Its equivalent in English typically has a sentential pronoun *it* as a subject. In the example (59) below, the antecedent is the copular subject, but there is no coordination. Moreover, the second token refers back to the first anaphor, thus characterising a chain.

(59) **A:** sim mas aqui fruta uma fruta né ?
gl: yes but here fruit a-FEM fruit not is ?
tr: yes but here fruit is any fruit isn't it ?

A: então quer dizer a senhora
gl: then want say-INF the-FEM lady

tr: then I mean you

A: pode ser a banana (CopFNP; ex_52; st; FtC;)
gl: may be the-FEM banana
tr: it may be a banana

A: pode ser a laranja (CopFNP; ex_52; st; FtCCh;)
gl: may be the-FEM orange
tr: it may be an orange

4.1.25 Copula-plus-adjective anaphor (CopAdj)

Similar to the type of anaphor above. However, the predicative is an adjective. Only one case was found in English as well.

(60) **C:** I don't recall saying that the deceased was drinking for three days I said on Sunday that she had drunk or was (CopAdj; ex_3; dthel; FtCCh;) drunk

Again this type of anaphor is a lot more common in Portuguese, with a large number of regular collocations. In the example (61) below, the anaphor does not refer to a definite antecedent. It is a collocation typically — but not necessarily — used as a discourse marker with the pragmatic function of requesting a sign of agreement or understanding.

(61) **A:** banana prata tem ... é de mais fácil digestão
gl: banana silver has ... is-EX of more easy digestion
tr: silver bananas have ... are easier to digest

A: que a banana d'água 'tá bom (CopAdj; NR; fdv; CK;) ?
gl: that the-FEM banana of-water-CONTR is-ST good ?
tr: than water bananas all right ?

4.1.26 Copula-plus-prepositional phrase anaphor (CopPP)

Ditto, but the predicative is a prepositional phrase.

(62) **A:** but I don't want them lost
B: certainly would be be (CopPP; ex_40; st; DK;) under lock and key before you leave the premises

As before, this type of anaphor is a great deal more frequent in Portuguese. One example is given below.

(63) **A:** e a senhora tem noção de como 'tá
gl: and the-FEM lady has notion of how is-ST

A: a pressão da senhora agora nos últimos dias
gl: the-FEM pressure of the-FEM lady now in-the-CONTR-MASCp last days

tr: and do you have any idea of how your pressure has been in the last days

B: não 'tava a a quinze por oito (CopPP; ex_16; st; FtC;)

Other predicatives can occur in anaphoric copular constructions in Portuguese. These types of anaphor are the **copula-plus-clause (CopCl)**, the **copula-plus-adverb (CopAdv)** and the **copula-plus-numeral (CopNUM)**. They are not as frequent as the preceding types, adding up to less than 1% of all cases identified in the annotated sample.

4.1.27 Reflexives (REF)

All occurrences of third-person reflexive pronouns plus first and second person tokens in verbatim reproduction of speech as in 4.1.3.

(64) **A:** but there's no indication there of who the writer is

B: no and we may well mother is mother uh coughs if she signs herself (REF; ex_64; thel; FtCCh;) mother

4.1.28 Reciprocals (REC)

All occurrences of *each other* and its variants. No occurrences were found in the samples of either language.

4.2 Type of antecedent

This property refers to the antecedent of a given anaphor token as identified by the analyst. It primarily concerns the explicit/implicit dichotomy. However, two other categories have been added for reasons which are explained below. A number is entered next to the code in order to identify the referent in a referent list which is kept in a separate file for each dialogue. This only applies to the first two categories, of course, as the two others classify antecedents which either do not exist or are too vague to be precisely defined. The code is entered in the second slot inside the brackets. Although the distinction between an implicit antecedent and an explicit antecedent is simple in most cases, it may involve fairly complicated decisions in some relatively rare occurrences.

4.2.1 Explicit (ex_)

The classification applies whenever the antecedent in question has been previously introduced in the dialogue or relates cataphorically to the anaphor in question. The antecedent may be a sentence, clause or chunk of discourse, provided it is clearly identifiable as realised in the text. For the sake of contrast, examples are given together with the examples for implicit antecedents below.

4.2.2 Implicit (im_)

The classification applies whenever the antecedent in question has not been previously introduced in the text nor relates cataphorically to the anaphor. It is often but not necessarily associated to nonpronominal noun phrases in relationships of hyponymy or superordination with noun phrases previously introduced. References which rely on world or shared knowledge for their resolution often belong in this category as well. Deictic references are seen as having an implicit antecedent if they have not been referred to by means of a nonpronominal noun phrase at any time previously in the dialogue. The example below contains cases of anaphora with both explicit and implicit antecedents. Note that the exchange opens one of the dialogues included in the sample.

(65) **A:** Mr Potter did you arrive about two o'clock on the Sunday the date (FNP; im_2; thel; SK;) the will (FNP; im_1; dt; SK;) was signed

B: yes (AdvR; ex_198; p_st; VMm;)

A: and did you go and see your mother (FNP; im_3; dthel; SK;) straight away

B: yes I did (OPT; ex_4; p_st; VMm;)

A: what was she (SP; ex_3; dthel; FtC;) then doing

B: she (SP; ex_3; dthel; FtCCh;) was having her (Pos; ex_3; dthel; FtCCh;) lunch

The first annotated token is *the Sunday the date the will was signed.* Strictly speaking, there are two tokens of anaphoric noun phrase, but in this case *the date the will was signed* is an apposition which uniquely characterises *the Sunday* in a way that resembles an adjectival relative clause. The token was then analysed as a single noun phrase. Having in mind that A's utterance is the first one in the dialogue, there is no previous mention of this discourse entity. However, the noun phrase is a definite description which assumes that the referent is already part of the body of knowledge shared by the participants and needs no introduction. The reference is thus anaphoric, although it is the first time that the entity is mentioned in the dialogue. Therefore, the antecedent is classified as **implicit**. A similar analysis applies to *will* and *your mother* below.

The reaction signal and the operator are responses to preceding questions. The antecedents have thus been previously introduced, leading to their classification as belonging to the **explicit** type. The sequence of references to B's mother by means of two personal pronouns and one determinative possessive are also classified as having an **explicit** antecedent because of the noun phrase which has introduced B's mother as a discourse entity. This pattern is common in dialogues where a body of shared knowledge is available. A noun phrase refers anaphorically to an implicit antecedent, and a number of references by means of pronouns follow this introductory token. There are of course many other contexts, both including noun phrases and other types of anaphor, in which antecedents are classified as implicit.

4.2.3 Nonreferential (NR)

The category refers to those occurrences which are not truly anaphoric, as there is no antecedent. The typical case are the occurrences of *it* in weather constructions and in some collocations, although *that* also appears as a nonreferential pronoun. It might be argued that these are not cases of anaphora at all, which is true. However, the pronouns involved are prototypical anaphors. Annotating them ensures that these cases will be unequivocally separated from the other true anaphors. An example is shown below.

(66) **A:** I suggest Mr Potter quite plainly that your mother telephoned the doctor and she was in a state of intoxication and it (SP; NR; fdv; CK;) was about three o'clock

As pointed out in 3.2, the assignment of referentiality to tokens of *it* and *that* is not such a simple matter, especially if the material analysed is taken from real-life dialogues. Thus, the generic standard adopted in the study found several occurrences of difficult classification. The generic standard was to consider all tokens of *it* which matched the definition of 'prop' *it* in [QGSL85], sections 6.17 and 10.26, as nonreferential. Occurrences of anticipatory *it*, as defined in [QGSL85], section 6.17, were classified as referential in all cases, that is, whether the occurrence was the subject of a cleft sentence ([QGSL85], section 18.25) or filled the position of an extraposed clausal constituent ([QGSL85], sections 18.33–35).

Cases of obligatory extraposition, such as the anticipatory *it* as a subject of the verb *seem*, were also considered as referential, as well as tokens of anticipatory *it* which refer to extraposed nonfinite clauses. All considerations of stilted or even impossible constructions generated by the replacement of an anticipatory *it* were left out of the classification standard. On the other hand, tokens of 'prop' *it* which might be arguably referential were considered to be nonreferential, such as those in which a temporal or locative phrase can be said to be an extraposed subject (see [QGSL85], section 10.26).

Tokens of *it* and *that* occurring in collocations which were found to be interpreted on the basis of their idiomatic value, as *that's it* or *I mean it*, were consistently classified as nonreferential, unless strong contextual evidence made this interpretation unacceptable. All tokens of nonreferential pronouns found in the sample appear in contexts which can be treated as collocations, including the tokens of 'prop' *it* as a subject pronoun, specified in [QGSL85] as typically 'expressions denoting time, distance, or atmospheric conditions'. However, constructions such as the one below were found in the corpus:

(67) **B:** you know I missed that one and I when I'd learnt about it
A: mm
B: er it (SP; NR; fdv; CK;) was too late
A: hm
B: I'd passed the deadline

The *too* preceding the time expression is often found in constructions which are followed by an infinitive clause. If this infinitive clause were actually explicit, the token would be analysed as an anticipatory *it* and, therefore, referential. On further analysis, the notion of an implicit extraposed subject was considered to be excessively dubious. These tokens were thus understood to be similar to those occuring in utterances like *it was late*. They were annotated accordingly, that is, as cases of nonreferential *it*. Possible interpretations in which the subordinate temporal clause would be analysed as a subject were also rejected on syntactic grounds, since these clauses have an adverbial function. This same interpretation led to the annotation of tokens as the one in example (68) below as nonreferential.

(68) **A:** you must take Joe Power's advice on this erm for the little that I know about the Ford Foundation and it is very little - um would have been that 'A' mightn't get anything you probably wouldn't get anything and 'B' it (SP; NR; fdv; CK;) would be a long time before you knew that you weren't getting anything

It is easy to think of a correspondence for the construction above which uses *it* + *take a long time*, followed by a to-infinitive clause with or without a subject. Therefore, the construction above might appear as *it would take a long time for you to know that you weren't get anything*. The standard adopted for the attribution of referentiality in this study analyses the correspondent construction as a case of anticipatory *it*, but it does not consider the *it* in construction such as the one in example (68) to be referential, as the subordinate clause has an adverbial function.

Tokens of *it* in rhetorical questions, such as the one in example (69) below, were analysed as nonreferential. Although the subsequent utterance might be interpreted as suggesting a correspondence with a that-clause, this is also thought to be an excessive "interference" of the analyst in what can be reasonably inferred from the actual corpus data.

(69) **A:** there's no Canadian money floating around is there hasn't Runnymede recently set up now what is it (SP; NR; fdv; CK;) hasn't some Canadian recently set up a foundation

The same problem of implicit extraposed subjects occurred in some highly standardised con-

structions such as the one below, which was also annotated as nonreferential. Although it is easy to think of a nonfinite clause, such as *talking to you*, as an implicit extraposed subject, assuming that this is the case was considered to be a disproportionately far-fetched conclusion in relation to the data at hand.

(70) **B:** thank you very much
A: it's (SP; NR; fdv; CK;) a pleasure

These examples suffice to characterise the standard used to decide on the referentiality of pronoun tokens. However, the analysis of anaphoric verbs in Portuguese raises further questions related to the notion of referentiality. As mentioned in the definition of anaphoric linking verbs, there are a number of collocations using verb forms with essential arguments missing which must be included in the analysis of anaphoric relations, just as all tokens of *it* and *that* were in the English sample. These verb forms may or may not perform pragamatic functions as well, and the complex interactions involving anaphoric relations, discourse markers and collocations will be discussed in another paper, as said before.

4.2.4 Discourse implicit (dim)

This category was created to account for relatively rare cases where the antecedent has to be built out of information given throughout the previous discourse, which amounts to a special form of implicitness. The pronoun *it* in C's second intervention below refers to a sequence of events in which the *doctor* gave a testimony the day before and was not challenged in his statement by the defendant's solicitors. However, the defendant is at the moment saying the *doctor* was wrong in his account of facts. The interpretation requires piecing together the whole chain of events, which are not described in an organised way, and understanding the consequences of having to call the doctor back for the courtroom procedures as an *unfortunate* development. The pronoun refers to the **problem** caused as a whole.

(71) **C:** so uh - and there it was - I think the doctor had better come back - this is a vital matter which had never been put to him - what d'you say Mr Hooker
D: if your lordship requires the doctor back
C: well it's (SP; dim; fdv; DK;) most unfortunate

Tokens of *it* as a subject pronoun may also occur in constructions which suggest a vague antecedent such as *the question*, *the situation*, or even *all that we talked about before*. These tokens were also annotated as cases of discourse-implicit antecedents.

(72) **B:** erm but it (SP; dim; fdv; DK;) was so curious because he said immediately without any hesitation he said have you tried erm the British Academy and I said yes

Although these definitions do not cover all possible nuances of referentiality, they provide effective guidelines for the vast majority of cases if used in combination with the collocation list.

4.3 Topical role of the antecedent

This property relates the antecedent to the topical roles discussed in Chapter 3. The systematic recording of a topical role for every antecedent is an attempt to quantify the relationship between topicality and anaphoric relations. The patterns uncovered by this variable are expected to shed light on the availability of certain elements for anaphoric reference of complex resolution. The categories include the topical roles mentioned in Chapter 3 and a few others needed for a thorough classification. The code for this property is entered in the third slot inside the brackets.

4.3.1 Discourse topic (dt)

The antecedent is the discourse topic as identified by the procedure in section 3.1.1.

4.3.2 Segment topic (st)

The antecedent is the segment topic as identified by the procedure in section 3.1.3.

4.3.3 Subsegment topic (sst)

The antecedent is the subsegment topic as identified by the procedure in section 3.1.3.

4.3.4 Discourse thematic elements (dthel)

These are elements of high saliency in a dialogue which are closely related to the discourse topic. Persons playing the role of agents, including the participants in the dialogue, are typical examples. They often appear as candidates for the role of discourse topic in frequency counts. Therefore, the procedure in section 3.1.1 will also provide the basic information to select discourse thematic elements. They should appear in the upper third of the frequency count for possible discourse topics. Also, the ratio used to assess even distribution should not be blatantly large.

4.3.5 Thematic elements (thel)

These are salient lexical items at the level of the segment, related to segment topics. They only occur within the scope of a segment or subsegment. The distinction between thematic elements and discourse thematic elements may often be blurred. One element may appear in two segments and then no more. This should not qualify it to be included in the group of discourse thematic elements, which should be kept as small as possible in order to be useful. Together with the frequency and distribution criteria specified for the category above, a thematic element should not be promoted to discourse thematic element if it appears in less than three nonadjacent nonresumptive segments.

4.3.6 Universal thematic elements (uthel)

The category contains anaphoric noun phrases which are universal — in the sense that their existence is assumed as given all the time — and thus quite invariably available for reference in spite of the current topic at any of the segmentation levels. An example of this sort of lexical item is *mother*.

4.3.7 Situational thematic element (sithel)

Anaphoric noun phrases such as *this country*, which are situational and also available for reference at any time during a dialogue.

4.3.8 Focusing device (fdv)

This category deals chiefly with nonreferential pronouns, although it may apply to cases of discourse-implicit antecedents. It does not classify the antecedent because there is no antecedent to classify. The focusing function is performed by the pronoun. As it does not actually refer, the role of this pronoun is to direct the processing, focusing the scope of the remaining part of the utterance. These phenomena are markedly different in Portuguese and it is important to annotate them for the purposes of contrastive analysis. An example is given below with the pronoun *that*.

(73) **A:** somebody took the tray out presumably .
B: er my wife took it out
A: and uh that's (De; NR; fdv; CK;) then about two fifteen
B: uh yes

4.3.9 Discourse chunks (p_)

Antecedents for the cases of sentential *it* and some occurrences of demonstratives are discourse chunks. These chunks are often but not always concerned with distinct items which have topical roles. The letter **p** for **predicate** is followed in the annotation by the underline character and one of the categories mentioned above, like in **p_st**, meaning **predicate of the segment topic**. One example is given below.

(74) **A:** I gather you've been at it for nine years

B: erm by golly that's (De; ex_3; p_st; CK;) true yes yes

4.4 Processing strategy

This property is an attempt to incorporate a psycholinguistic element into the annotation. The code is placed in the fourth slot inside the brackets. It can be defined as a guess as to the most important form of knowledge for the successful resolution of the anaphora in question. Its purpose is to enrich the classification by classifying the casess of anaphora according to processing, uncovering distinctions which might remain unnoticed if only the type of anaphor were to be specified. It is hoped that the classification according to the categories in this property could guide anaphora resolution in a computer system.

4.4.1 Shared knowledge (SK)

Dialogues evolve on the basis of presupositions concerning what participants already know and do not have to be told about. The information is used to interpret anaphors such as the noun phrase *the thesis* in the example below. Both participants know that *B* is writing a thesis at the moment, therefore, the noun phrase does not have to be introduced before it is used as an anaphoric noun phrase.

(75) **A:** how's the thesis (FNP; im_1; dt; SK;) going

B: uh I'm typing it up now typing up the final copy (FNP; im_1; dt; WK;)

4.4.2 World knowledge (WK)

The second nonpronominal anaphoric noun phrase in (75) above (*the final copy*) depends on knowledge about theses in general. Although restricted to academic life, this kind of information is better classified as **world knowledge** because, unlike the knowledge used to process the first noun phrase, it is relatively invariant over time and place. There should be a *final copy* to every *thesis*, and that is the information needed to process the reference and find the implicit antecedent (B's thesis).

4.4.3 Lexical signalling

This type of anaphoric noun phrase is resolved using primarily dictionary-like knowledge. In the example below, the word *monies* relates to *finances* not by repetition or modification of the verbatim form of the antecedent but through a connection based quite exclusively on the lexical content of the anaphoric noun phrase.

(76) **B:** and uh - you know my own personal finances are

A: well sure

B: it's just out

A: but you have applied er for monies (FNP; im_12; st; LS;) I keep hearing wherever I go

Distinguishing these three types may not be easy in some cases. In fact, the three categories classify a continuum of knowledge, which begins with **shared knowledge** at one end and ends with **lexical signalling** at the other. In spite of the arbitrariness involved in some decisions, the separation of the processing in three distinct strategies was considered useful.

4.4.4 Lexical repetition (LR)

This category classifies anaphoric reference in which the simple repetition of a lexical item is the clue to establish the link between the anaphor and the antecedent, as in the following example.

(77)
B: and erm I don't know why Oxford turned it down I don't really know why Cambridge turned it down I mean it's got to be done by a university press because it's not going to be a remunerative thing you know it

A: mm mm

B: well it's not a best-seller

A: obviously the university presses (FNP; ex_50; sst; LR;) are in recent years very reluctant to undertake big schemes of this kind

4.4.5 Modified antecedent (AM)

In the example above, the only difference between the two noun phrases is the determiner used. For the purposes of the research, a change of determiner, especially the variation from indefinite article to definite article, as in the example above, is still seen as simple repetition. However, some cases of anaphora involve a more significant alteration of the antecedent. In the example below, the anaphoric noun phrase has a different head from the antecedent, although the change is not enough to disrupt the link in the anaphoric relation.

(78)
A: where have you been working

B: I've been working as a research assistant for Professor Leegate on the collected notebooks of Etheridge

A: oh yes

B: erm I did most of the documentation for volume three

A: Professor Leegate

B: yes

A: is this the Canadian girl

B: Caroline Caroline

A: yes yes

B: I did most of the documentation research (FNP; ex_7; sst; AM;) for volume three for her

4.4.6 First candidate search (FtC)

This is the typical case of pronoun resolution described in [Hob86] as accomplished by a 'naive algorithm'. An adaptation of the algorithm for spoken language would specify the same kind of processing, as the strategy consists of a search for the antecedent on the basis of syntactic

knowledge — basically the notion of command — and agreement constraints. The first suitable candidate found in such a search is the antecedent for the anaphor. An example is given below:

(79) **A:** how's the thesis going
B: uh I'm typing it (OP; ex_1; dt; FtC;) up now

4.4.7 First candidate chain (FtCCh)

This is the kind of anaphoric reference that has an anaphor as an antecedent. This anaphor antecedent can be identified by means of a putative adapted version of Hobbs' algorithm. The last anaphor in the chain will be linked to the common antecedent, as in the example below:

(80) **B:** and I went down this morning to talk to the American Embassy on the off chance that the State Department might be you know able to finance a bit of travelling in the States and they can't they've (SP; ex_13; st; FtCCh;) got priority on vice-chancellors and uh English schoolteachers

It is important to note that this processing strategy classifies pronominal anaphors which have another pronominal anaphor as the first candidate for antecedent. The preceding anaphor is the correct antecedent because it either links to a correct common referent or to another anaphor in a chain which ultimately links to a correct common referent. The classification implies nothing about the processing strategy to resolve the antecedent anaphor or about the anaphor that starts the chain. The anaphor that starts the chain does not have to be resolved by means of a first-candidate strategy, although this is true in a large number of cases. A pronoun can start a chain even if the antecedent is only identifiable by means of deixis or discourse knowledge (see below).

As the Portuguese referring system relies on argument structure rather than explicit markers of anaphoric relations, such as pronouns, the notion of first candidate search had to be adapted accordingly. The assignment of first-candidate strategy to tokens of anaphoric verbs was then changed to mean the first entity with the same syntactic function. If this entity was an omitted argument of a preceding occurrence of anaphoric verb, the processing strategy was assigned the value of first-candidate chain. One example of these occurrences was shown in the definition of anaphoric verbs above (example (44)). Another one is shown below.

(81) **A:** a senhora perdeu trezentos gramas isso aí é assim
gl: the lady lost-PAST3rds three-hundred grams this there is like
tr: you lost three hundred grams this is like

B: perdi (VerbAn; ex_143; thel; FtC;) ?
gl: lost-PAST1sts
tr: did I

A: perdeu (VerbAn; ex_143; thel; FtCCh;)
gl: lost-PAST3rds
tr: you did

In **B**'s first utterance, the verb form *perdi* appears stripped of both subject and object. The omission of the subject is not considered to be an anaphoric reference, as the first-person morphology is inequivocal, but the omitted object has to be retrieved for the question to be interpreted. As the first verb form to be found in a search backwards is the third person inflection of the same verb in the same tense, the explicit direct object of this verb form is the antecedent of the sub-

sequent anaphoric verb. The same direct object is also omitted in **A**'s response to **B**'s question, characterising the first-candidate chain.

4.4.8 Verbatim memory (VMm)

Anaphora resolution may rely on literally recalling the exact terms of the antecedent as it was uttered. This seems to be particularly important for the types of anaphor which involve ellipses of sentences and verb phrases. As speech is evanescent, such strategy demands recency as a precondition for the verbatim retrieval of utterances.

(82) **A:** I gather you've been at it for nine years
B: erm by golly that's true yes yes it's not a long time of course in the uh in this sort of work you know
A: well no but it's quite a long time by any standards
B: yes suppose so (SoAn; ex_6; p_st; VMm;)

4.4.9 Parallel (Pl)

The identification of the antecedent may rely on processing which involves parallelism of syntactic structures. Thus, in the example below, the pronouns *he* and *him* can only be resolved by using the information which defines the syntactic functions of antecedents in the previous move. These syntactic functions are retained in the subsequent move.

(83) **A:** well of course a stockbroker doesn't do that he merely takes on Mr Y as a client and he (SP; ex_220; dt; Pl;) does his best for him (OP; ex_149; dthel; Pl;)

If the parallel strategy were to be applied as described above to Portuguese, virtually all anaphoric verbs and linking verbs with an explicit antecedent would be classified as relying on parallel strategy, as anaphora resolution relies on argument structure. Bearing in mind the adaptation described above, the parallel strategy is assigned, in Portuguese, to cases which demand overriding the information in the first verb form found in a backward search on the basis of syntactic information. One example is given below.

(84) **A:** você tem gases , costuma ter , assim ?
gl: you have gases , wont-PRES3s to have , so ?
tr: do you usually have trapped air ?

A: porque , costuma dar , né , uma uma um incômodo
gl: because , wont to give , not-is-CONTR , a a a discomfort
tr: because it usually causes a discomfort

B: diz que dá
gl: says that gives
tr: it is said it does

A: é , dor mesmo , dá dor mesmo
gl: is , pain same , gives pain same
tr: yes , real pain . it causes real pain

B: é . diz que dá (VerbAn; ex_110; sst; Pl;) muita dor até

gl: is , says that gives , much pain even

tr: yes , it is said it does , a lot of pain

The verb *dar* appears for the first time in **A**'s final move within the first turn, as the head of the verb phrase *costuma dar*, in which it is linked with the third person singular verb form *costuma*, present tense of the catenative verb *costumar*, indicative mode. It appears for the second time with both arguments omitted in a nominative subordinate clause in **B**'s subsequent utterance. It then appears again as the main verb in an utterance with a new but semantically related object and the subject still omitted, followed by a new token identical to the one which appeared in the turn before the last. The two tokens in subordinate clauses demand a much stronger parallelism effect to be resolved, which requires bypassing the main verb and retrieving the antecedents in the previous turn.

4.4.10 Discourse knowledge (DK)

This category classifies anaphora cases which require full processing of discourse for its resolution. In the example below, the implicit antecedent *number of cassettes* can only be identified by means of a gradual buildup involving all levels of information in a joint effect.

(85) **A:** what they've done is I think you know several thousand of the book this paperback but only three hundred of the cassettes er so there's that disparity uh in the marketing so that they're they're quite willing to conceive of of this kind of disparity but scaled down you know it (SP; im_57; sst; DK;) might be a hundred

The implicit antecedent may relate to an already explicit one in a form of superordination which is assumed as easily inferred by the listener. Thus, in example (86) below, the last two tokens of *it* refer to *theses in general* and not to *B's thesis*.

(86) **A:** how's the thesis going

B: uh I'm typing it up, typing up the final copy

A: hm uh when are you submitting it

B: erh - well it it would have been

A: next term

B: this autumn but er I had to go to work this winter and that really

A: but if you're typing it up now er why can't

B: it's going so slowly though you know it's this it's these awful these awful symbols .

A: mm

B: you know it's a combination of of the phonetic alphabet .

A: mm

B: plus the reformed spelling you know how it (SP; im_2; dt; DK;) is uh you can't rush it (OP; im_2; dt; DK;)

The category also covers processing strategies which handle hard cases, when an intervening acceptable antecedent has to be ignored in favour of the intended one, as in the occurrences below.

(87) **A:** so again Mrs Kay is wrong in this (De; ex_106; p_st; DK;) let's just read on she (SP; ex_3; dthel; DK;) did not say I'm going to do it (DPA; ex_88; p_dt; DK;) again as I don't want there to be

any trouble is she (SP; ex_221; dthel; DK;) wrong in that (De; ex_108; p_st; DK;)

B: yes I would take it that she didn't remember it (OP (cataph); ex_109; p_dthel; DK;)

As previously spelled out, this strategy is assigned to Portuguese tokens of anaphoric verbs which demand the same sort of bypassing operation for the identification of the correct antecedent. Anaphors with discourse-chunk antecedents of complex identification, as the one in example (88), are also included in this category.

(88) **A:** but I've always been told that diarists are crazy as well

B: um well there may be of course something in this (De; ex_19; p_st; DK;) but

4.4.11 Set member selection (SetMb)

Anaphora occurrences assigned to this category rely on processing involving a previously mentioned set of objects from which one specific member is selected by the anaphor. The strategy is often but not exclusively associated with one-anaphoras.

(89) **B:** we replace all the proper names including place names

A: mm yes mhm

B: by fictitious ones (One_an; ex_33; thel; SetMb;)

4.4.12 Set creation (SetCr)

The anaphoric reference may also create a set of objects to refer to various objects previously mentioned separately.

(90) **B:** if you want to have philosophy and uh mathematics as your your two possible subjects (FNP; ex_26; thel; SetCr;) as an undergraduate then you can do those

A: oh no

4.4.13 Collocations (CK)

Anaphoric pronouns appear in collocations such as *that's it* or *to put it mildly*. The interpretation of these tokens differs in a regular way from the expected first-candidate strategies, taking a predictable resolution path according to the collocation in which they appear. The collocations for the English anaphors were listed in order to allow recognition of these occurrences in association with their respective regular resolution paths. A similar list was collected for the Portuguese anaphors with the same sort of regular patterns of recognition and resolution. The Portuguese collocations, as expected, contain verb forms more frequently than pronouns. Characteristics of the entries in the list are not discussed in this paper, but see [Roc97]. See also [Roc98] for the complete list of collocations in both languages.

(91) **B:** the bibliography has gone about as far as I can take it on my own that (De; ex_10; p_st; CK;) is to say er in order to complete it I will have to visit the major resources in the United States and uh several in Europe

4.4.14 Secondary reference (ScRf)

First and second person pronouns are not normally anaphoric. However, they may refer anaphorically when speech is reported verbatim. This category accounts for such cases. The same definition applies for first and second person verb forms in Portuguese.

(92) **A:** and I said if this is what you (SP; ex_8; dthel; ScRf;) want I would put the maximum pressure upon somebody like Derek Brainback to do it

4.4.15 Deixis (Dx)

Anaphora cases are assigned to this category when the identification of the antecedent depends on information available in the physical environment where the dialogue occurs.

(93) **A:** here is your copy ((Pos) FNP; im_1; st; Dx;) of the revised version and I'll stick that (De; im_3; st; Dx;) in Gavin's pigeonhole

4.5 Borderline cases

The fact that distinctions are not so easily established is a constant problem for taxonomies dealing with natural languages, as the analysis relies on categories rather than scores or rankings for the assignment of cases to classes. This difficulty was made more punishing by the fact that this investigation developed the taxonomy along with the analysis of corpus data. Although a number of intuitions and beliefs were transformed into an initial classification, so as to allow the annotation work to start, the categories were shaped by the routine of analytical work, an unavoidable consequence of the exploration into terra incognita which the investigation required. Borderline cases were repeatedly reannotated as a result of changes in orientation.

The way to deal with borderline cases suggested in [Edw92] is the technique of double coding. Whenever a firm decision cannot be made, two possible solutions are entered. The problem with this kind of solution is that, for the purposes of statistical analysis, double coding is tantamount to creating a new category. If a significant number of cases require the same sort of double coding, the introduction of a new category is likely to be the best solution. If distinct combinations of double coding are used too often, the number of categories may grow beyond a reasonably manageable amount. As the classification used in this investigation can be said to be fairly detailed, double coding was avoided. Although some cases were provisionally double-coded, pending a last-minute decision, this decision was made before the statistical analysis began.

Some typical problems were solved in the simplest way possible. For instance, some explicit antecedents may be so vague as to challenge the notion of explicitness, as occurs when they are introduced by a noun phrase of low semantic content, such as *this sort of thing*. For the purposes of this research, whenever an antecedent had been introduced by a nonpronominal noun phrase, it was considered to be **explicit**. Whenever it was introduced by a pronoun, for instance by means of deictic reference, and then referred to in a chain, it was considered to be **implicit**.

The classification of antecedents according to their topical roles inevitably leads to overlapping of functions. Thus, a discourse entity which is salient enough to be included in the list of **discourse thematic elements** may become a **segment topic** or **subsegment topic** at some point throughout the discourse. Whenever there was reference to an entity in a situation of this kind, preference was given to local roles, as segment and subsegment topic, over global roles, such as discourse thematic elements. The problem does not occur with the single element chosen as the **discourse topic**, which was never chosen as a local topic.

The very notion of a processing strategy clearly involves a degree of abstraction, in the sense that the classification typically refers to the most important form of knowledge used for the res-

olution of the anaphor, seldom to a single form of knowledge, as several are needed. Thus, an element of **verbatim memory** is invariably involved in a resolution based on **parallel**, and, to a certain extent, the reverse is also true. However, in the examples given above, the anaphor resolved by means of **verbatim memory** requires the literal form of the previous move for its resolution. On the other hand, the syntactic function of the words do not play a direct role in the resolution. Contrastively, the example of **parallel** can only be handled by using the syntactic function of the antecedents as a guide for the resolution. The retention of the literal form in itself is not enough to accomplish the identification of the correct antecedents. Thus, the classification according to processing strategy aims at the crucial knowledge rather than at all knowledge involved in the resolution.

Chapter 5

Conclusion

This paper was intended as a relatively succinct description of the annotation scheme aimed at readers who are primarily interested in the annotation proper. The annotation scheme was created as part of a larger project which involved its use to analyse a large number of anaphora cases in English and Portuguese. Statistical techniques were subsequently used to further explore the results concerning each one of the languages. Ultimately, a systematic description of recognition and resolution patterns in both languages, named the antecedent-likelihood theory, was organised and tested. These results were then used in a contrastive analysis of anaphoric relations in dialogues in English and Portuguese. The project eventually became a DPhil thesis ([Roc98]).

On the other hand, the properties included in the annotation had to be sufficiently discussed in order to ensure that the conceptual framework behind the annotation was properly understood. Thus, it was necessary to make the paper somewhat lengthy. Nevertheless, a number of issues, such as a thorough contrastive analysis of the anaphoric relations in the two languages, as well as the interaction between anaphora, collocations and discourse markers, were barely discussed. Work quoted throughout the paper can be sought by those who may have an interest in these developments.

Finally, the annotation scheme underwent many changes during the annotation of the samples. Although it seems reasonable to assume that it has now reached a relatively stable form, it can be used in a variety of ways, according to specific demands. The crucial feature of the annotation is the choice of properties included. Changes which do not eliminate any of the four properties — grouping of categories, for instance, being an obvious one — are possible and may result in a more effective version of the scheme for different research purposes. In fact, umbrella categories were created for the statistical analysis carried out in the DPhil thesis project (see [Roc98]). The scheme is thus thought of as a useful starting point. Future developments are planned, involving the conversion of the scheme to a more universally known code, such as SGML. In time, it is hoped that the scheme will be useful for the research community involved in corpus linguistics and related fields.

Bibliography

[All87] James Allen. *Natural language understanding*. Benjamin and Cummings, Menlo Park, CA, 1987.

[Bib92] Douglas Biber. Using computer-based text corpora to analyse the referential strategies of spoken and written texts. In *Directions in corpus linguistics*, pages 215–252. Mouton de Gruyter, Berlin, 1992.

[Bos83] P. Bosch. *Agreement and anaphora*. Academic Press, New York, 1983.

[BY83] Gillian Brown and George Yule. *Discourse analysis*. Cambridge University Press, Cambridge, 1983.

[CC85] Celso Cunha and Lindley Cintra. *Nova Gramática do Português Contemporâneo*. Nova Fronteira, Rio de Janeiro, 2nd edition, 1985.

[Edw92] Jane Edwards. Design principles in the transcription of spoken discourse. In Jan Svartvik, editor, *Directions in corpus linguistics*, pages 139–144, Berlin, 1992. Mouton de Gruyter.

[FH92] Gill Francis and Susan Hunston. Analysing everyday conversation. In R.Coulthard, editor, *Advances in Spoken Discourse Analysis*. Routledge, London, 1992.

[Fli92] Steve Fligelstone. Developing a scheme for annotating text to show anaphoric relations. In G. Leitner, editor, *New directions in English language corpora: methodology, results, software development*, Topics in English Linguistics no. 9, pages 153–170. Mouton de Gruyter, Berlin and New York, 1992.

[Fox87] Barbara Fox. *Discourse structure and anaphora*. Cambridge University Press, Cambridge, 1987.

[GJW95] Barbara Grosz, Aravind Joshi, and Scott Weinstein. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225, 1995.

[GS86] Barbara Grosz and Candace Sidner. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

[HH76] M.A.K. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

[Hob86] 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.

[Hoe91] Michael Hoey. *Patterns of lexis in text*. Oxford University Press, Oxford, 1991.

[Kje91] Goran Kjellmer. A mint of phrases. In K. Aijmer and B. Altenberg, editors, *English corpus linguistics: studies in honour of Jan Svartvik*. Longman, Harlow, 1991.

[Lan69] R. Langacker. On pronominalization and the chain of command. In D. Reibel and S. Schane, editors, *Modern studies in English*, pages 160–186. Prentice-Hall, Englewood Cliffs, 1969.

[QG73] Randolph Quirk and Sidney Greenbaum. *A university grammar of English*. Longman, 1973.

[QGSL85] Randolph Quirk, Sidney Greenbaum, Jan Svartvik, and Geoffrey Leech. *A comprehensive grammar of the English language*. Longman, 1985.

[Rei83] Tanya Reinhart. *Anaphora and semantic interpretation*. Croom Helm, London, 1983.

[Roc97] Marco Rocha. Supporting anaphor resolution in dialogues with a corpus-based probabilistic approach. In Ruslan Mitkov and Branimir Boguraev, editors, *Operational factors in practical, robust anaphora resolution for unrestricted texts*, pages 54–61. Association for Computational Linguistics, July 1997.

[Roc98] Marco Rocha. *A corpus-based study of anaphora in dialogues in English and Portuguese*. PhD thesis, School of Cognitive and Computing Sciences, University of Sussex, 1998.

[Sam87] Geoffrey Sampson. MT: A nonconformist's view of the state of the art. In Margaret King, editor, *Machine translation today: the state of the art*, pages 91–108. Edinburgh University Press, Edinburgh, 1987.

[SC92] John Sinclair and Malcolm Coulthard. Towards an analysis of discourse. In R.M. Coulthard, editor, *Advances in Spoken Discourse Analysis*. Routledge, London, 1992.

[Sin92] John Sinclair. Priorities in discourse analysis. In R. Coulthard, editor, *Advances in Spoken Discourse Analysis*. Routledge, London, 1992.

[Sin93] 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.

[SSJ74] H. Sacks, E. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735, 1974.

[vDK83] T. van Dijk and W. Kintsch. *Strategies of discourse comprehension*. Academic Press, New York, 1983.

[Web79] Bonnie Lynn Webber. *A Formal Approach to Discourse Anaphora*. Garland, New York, 1979.

Appendix A

Conventions

A.1 General conventions

The symbols below are used in both versions of the AL theory and of the collocation list. They also appear at examples and glosses in the text. COL - collocation
X-*verb* - any form of the verb
NP - noun phrase
ADJ - adjective
modif - modifier
Compl - complement of verb
LV - linking verb
PP - prepositional phrase
Obj - object
ObjP - object pronoun
NUM - numeral
AdvP - adverb of place
NF-clause - non-finite clause
IndArt - indefinite article
Pos - possessive
DET - determiner
SubjC - subject complement
De - demonstrative
Art - article
SP - subject pronoun
Subj - subject
PastP - past participle
INF - infinitive

A.2 Conventions used in the glosses

Portuguese examples appear with a gloss (marked **gl:** underneath, followed by a translation (marked **tr:**). Glosses only include a morpheme-by-morpheme account when thought necessary. Translations are included as a guidance only, as it is sometimes difficult to find precise solutions in English for colloquial expressions in Rio de Janeiro Portuguese. In cases in which the translation would be identical to the gloss, the translation is left out.

Several conventions used in the glosses indicate morphological features. However, morpheme-

by-morpheme analyses are only included when seen as crucial for the point being made. For instance, if the past tense form *passei* is preceded by the first person pronoun *eu* in the Portuguese speech, the gloss may include simply *I passed* without the person, tense and mode specifications, unless where the inflection is essential to the discussion.

In some cases, extra information about the examples is provided between brackets next to the words concerned. Thus, *bacurau*, a Brazilian bird, appears in the glosses as *bacurau (a bird)*. Any other relevant information may be included using this convention. Symbols listed in the previous session are also used in the glosses.

A.2.1 Verb tenses

PRES - present PAST - past INF - infinitive FSU - future subjunctive IMP - imperfect past

A.2.2 Verbal persons

1st - first person
2nd - second person
3rd - third person

A.2.3 Number

s - singular
p - plural

A.2.4 Gender

MASC - masculine
FEM - feminine

A.2.5 Miscellaneous

CONTR - contraction (preceded by contracted words linked by a hyphen)
DIM - diminutive
X-be-EX - existential *be*
X-be-ST - stative *be*
there-N - adverb of place meaning close to both speakers
there-F - adverb of place meaning away from both speakers
there-NS - adverb of place meaning away from speaker but close to hearer

Appendix B

Quick reference for code in the annotation scheme

B.1 Types of anaphor

Nonpronominal noun phrase	FNP	Do-phrase anaphora	DPA
Anaphoric adjective	AdjAn	Linking verb	LV
Subject pronoun	SP	Operator	OPT
Object pronoun	OP	Anaphoric Verb	VerbAn
Demonstrative	De	Copula-plus-noun phrase anaphor	CopFNP
Determinative possessive	Pos	Copula-plus-adjective anaphor	CopAdj
Independent possessive	PPos	Copula-plus-prepositional phrase anaphor	CopPP
One-anaphora	One_an	Anaphoric non-finite clause	NFClAn
Reciprocal	REC	Anaphoric that-clause	TCLAn
Reflexives	REF	Adverb of time	AdvT
Adverb of place	AdvP	Adverb of manner	AdvM
Numeral	NUM	Adverb of frequency	AdvF
Indefinite pronoun	IP	Copula-plus-adverb anaphor	CopAdv
Wh-word	WHT	Adverb of intensity	AdvI
Prepositional phrase	PP	Copula-plus-clause anaphor	CopCl
Reaction signal	AdvR	Copula-plus-numeral anaphor	CopNUM
So-anaphora	SoAn	Adverb of exclusion	AdvE

B.2 Types of antecedent

explicit	ex_
implicit	im_
nonreferential	NR
discourse implicit	dim

B.3 Topical roles of the antecedent

segment topic	st
discourse topic	dt
subsegment topic	sst
thematic element	thel
focusing device	fdv
discourse thematic element	dthel
universal thematic element	uthel
situational thematic element	sithel
predicate of segment topic	p_st
predicate of discourse topic	p_dt
predicate of subsegment topic	p_sst
predicate of thematic element	p_thel
predicate of focusing device	p_fdv
predicate of discourse thematic element	p_dthel
predicate of universal thematic element	p_uthel
predicate of situational thematic element	p_sithel

B.4 Processing strategies

Shared knowledge	SK
World knowledge	WK
Lexical repetition	LR
Lexical signalling	LS
Set member	SetMb
Set creation	SetCr
Collocational knowledge	CK
First candidate	FtC
First candidate chain	FtCCh
Modified antecedent	AM
Verbatim memory	VMm
Secondary reference	ScRf
Parallel	Pl
Discourse knowledge	DK
Deixis	Dx