

Dialogue Acts, Synchronising Units and Anaphora Resolution

Miriam Eckert

University of Pennsylvania

Michael Strube

European Media Lab

Abstract

In this paper, we present the results of a corpus analysis, and a model of anaphora resolution in spontaneous spoken dialogues in the form of an algorithm. The main finding of our corpus analysis is that less than half the pronouns and demonstratives have NP antecedents in the preceding text. 22% have sentential antecedents and the remainder have no identifiable linguistic antecedents. As part of the corpus analysis we present the results of inter-annotator agreement tests. These were carried out for marking anaphor types and their antecedents, and for segmenting the dialogues into dialogue acts. The results of the inter-annotator agreement tests indicate that our classification method is reliable and that the annotated dialogues can be used as a standard against which to measure the performance of the resolution algorithm. The algorithm, based on Strube (1998), is capable of classifying pronouns and demonstratives, and co-indexing anaphors with NP and sentential antecedents. The domain from which potential antecedents for both individual and discourse-deictic anaphors can be elicited is defined in terms of dialogue acts. The dialogue segmentation method uses dialogue acts to form *Synchronising Units*, which reflect the achievement of *common ground* (Stalnaker, 1974; 1979). We show that predicate information, NP form and dialogue structure can be successfully used in the resolution process.

1 Introduction

In this paper, we present a model for the resolution of pronouns and demonstratives in spontaneous spoken dialogue. In the semantic, syntactic and psycholinguistic literature, work on anaphora has concentrated primarily on the analysis of pronouns and definite NPs with NP-antecedents. This is considered to be the “normal” type of anaphoric reference. Our corpus study reveals that in actual language use, this type of anaphoric reference accounts for less than half of the occurrences of pronouns and demonstratives (45%). An additional 22% are anaphors with sentential and VP-antecedents. Although this type has been studied previously (Webber (1991) and, particularly, Asher (1993) provide extensive theoretical accounts), it seems that its frequency and therefore importance has been largely underestimated. Rather surprisingly also, the remaining third of all pronouns do not have identifiable linguistic antecedents of any kind. These are pronouns which refer to inferrable entities and those which refer to a vaguely defined general discourse topic. These findings indicate that an important function of pronouns, aside from anaphoric reference, is that they allow the speaker to leave certain referents underspecified. In spontaneous spoken language it is simply not necessary for the participants to be able to unambiguously identify a specific referent at all times. If they fail to understand an utterance and consider avoidance of misunderstanding to be important, they can immediately request clarification – an option not available in the written medium. Furthermore, the optional use of vague pronouns greatly facilitates the task of the speaker in on-line language production.

We present a model which shows how pronouns and demonstratives can be classified and, if appropriate, co-indexed with the correct antecedents. The model makes use of the surface form of the anaphor, its predicative context and the structure of the discourse. It also presents a basis for further empirical evaluations of theoretical issues in anaphora resolution. Furthermore, we believe that it provides an important starting point for spoken language resolution algorithms in the field of computational linguistics, which have so far exclusively dealt with anaphora in written texts.

In computational linguistics, most anaphora resolution algorithms are designed to deal with the predominant type of anaphoric reference found in written texts, which involves the co-indexing relations between anaphors and NP-antecedents. Aside from the different types of anaphors found in spoken language, the structure of dialogues is less clear than the structure of written texts, with lack of punctuation, paragraphs and syntactically complete clauses making it difficult to formally define the domain for potential antecedents. For these reasons, applying existing anaphora resolution algorithms to dialogues would result in a poor performance.

Our model is presented in the form of a major extension of the anaphora resolution algorithm described in Strube (1998). The Strube (1998) algorithm consists of an ordered list of salient discourse entities (S-List), which provides preferences for the antecedents of pronouns. The main characteristic of the algorithm is that preferences for intra- and intersentential pronouns are dealt with in a unified manner, as the update of the S-List and the anaphora resolution are performed incrementally. Essential to the success of the algorithm presented in this paper is the interaction between the identification and resolution of different types of anaphors and the deter-

mination of the domain of possible antecedents. We use dialogue act units (derived from speech acts) to provide the structure necessary for the determination of the antecedent domain and also use them to function as antecedents for anaphors with sentential antecedents.

The paper is structured as follows: Section 2 describes the theoretical issues which are important for our analysis and which have partly been incorporated into the algorithm. Section 3 describes the spoken language corpus used for our empirical analysis of anaphor types and for testing the algorithm. Section 4 gives an overview of our classification system for the different types of pronouns and demonstratives we identified in the spoken dialogues. Section 5 describes how we use dialogue acts to model the establishment of common ground and to define the domain of possible antecedents for the anaphors. Our resolution algorithm is presented in Section 6. Section 7 gives the results of the empirical analysis. This consists of two parts: first, we evaluated the classification system in terms of inter-annotator agreement. We deemed this step necessary in order to verify the consistency of our classification. Second, we evaluated the algorithm by applying it to the hand annotated dialogues. Sections 8 and 9 provide comparisons to related work, suggest future additions and applications of our model, and present the conclusions.

2 Theoretical Issues

In this section, we present some of the issues in theoretical linguistics, which we consider to be important for the process of anaphora resolution in spoken dialogue.

The value of these issues has so far been expressed in theoretical terms. We consider one of the contributions of our resolution algorithm to be that it opens the possibility of testing their value empirically.

2.1 Reference and the Discourse Model

We assume that a conversation has associated with it a model of the discourse, which is distinct from both the real world and from the syntactic representation of the discourse. Such models have frequently been described in the literature, e.g. *common (back)ground* (Stalnaker, 1974; 1979), *discourse model* (Webber, 1979), *files* (Heim, 1982), *attentional state* (Grosz & Sidner, 1986), *DRSs* (Kamp & Reyle, 1993). These proposed models differ in a number of important ways, such as whether they are said to exist at the semantic level (*files*, *DRSs*), the pragmatic level (Stalnaker’s *common ground*) or the discourse level, (Webber’s *discourse model*, Grosz & Sidner’s *attentional state*). Also, some models are proposed to be properties of the conversational participants (Stalnaker’s *pragmatic presuppositions* constituting the common ground), whilst others are properties of the discourse itself (*DRSs*, *attentional state*).

What these versions of the discourse model have in common is that they are assumed to contain representations of the objects that have been referred to in the discourse, known as the *discourse referents* (Karttunen, 1976), *file cards* (Heim, 1982) or *discourse entities* (Webber, 1979; Kamp & Reyle, 1993). The discourse model also contains the attributes of the discourse entities and the relations holding between them but, for the moment, we will focus only on the entities introduced by NPs in

the discourse. The discourse model contains representations of the entities that are salient to both participants at a given point in the discourse because they have been referred to in the previous discourse. Using terminology from Stalnaker (1979) and Clark & Schaefer (1989), we will call the part of the model containing representations of these entities *common ground*.

The update of the discourse model has been the subject of considerable debate. One issue is the question of when and how entities are entered into the common ground. Seeing as conversations involve more than one participant, merely uttering a sentence does not mean that the entities referred to have been entered into the common ground. It is possible, for example, for one speaker to ignore the utterance of another. Conversational participants have a number of ways in which to signal understanding of an utterance, including nods of the head, relevant further contributions to the discourse, and simple backchannels (e.g. *uhu*, *yeah*, *mmhm*). In our model, if an utterance is not acknowledged by the other participant, its discourse entities are not retained in the common ground. This issue is explained in more detail in Section 5.

In addition to this issue, there has been disagreement concerning the influence of NP form on update, that is, whether indefinite NPs, definite NPs and pronouns serve to update the discourse model in the same way or whether different mechanisms need to be postulated. In Russell's view (Russell, 1905), indefinite NPs are not referring expressions, but rather function much like existential operators, by declaring that the set of entities described by the NP is not null. This view was subsequently challenged because it does not explain the capacity of indefinite NPs to function as antecedents of anaphoric pronouns (Grice, 1975; Kripke, 1979; Lewis, 1979). In Heim's file change

semantics (Heim, 1982), the approach is taken that indefinite NPs introduce new entities (*file cards*) to the discourse model, whereas definite NPs make use of familiar ones.

A concern with making a categorical distinction between definites as NPs referring to given entities, and indefinites as NPs referring to new entities, is that there are many counterexamples, in which definites are used to refer to discourse-new entities (Prince, 1981). In fact, recent empirical research has indicated that the numbers are by no means negligible. Poesio & Vieira (1998) show that in their corpus only 50% of definites are discourse-new. The reason is that, as noted by Prince (1981) and Prince (1992), the status of entities is far more complex than can be determined by the distinction *given – new*.

Prince adds the distinction *hearer-old – hearer-new* to the *discourse-old – discourse-new* factor. *Discourse-old/new* describes the information status of an entity with respect to the discourse. *Hearer-old/new* describes the status with respect to the hearer. A definite NP such as *your father*, for example, can be discourse-new if its referent has not been mentioned before, but *hearer-old* because it is clearly familiar to the addressee. In addition to these distinctions, there are also *inferrable* entities. These are hearer-new, discourse-new, but “depend upon beliefs assumed to be hearer-old, where these beliefs crucially involve some trigger entity” (Prince, 1992). A trigger entity can be the referent of a previously mentioned NP, e.g. *a house*, which once established in the discourse, allows one to refer to expected and related entities such as *the front door* with a definite NP.

The discourse model is not intended to reflect which entities are familiar to the

hearer but rather which entities are *salient* at that point in the discourse. We therefore assume that indefinite and definite NPs can add entities to the discourse model because they can both cause a referent to become salient in the discourse. The category inferrable is only accounted for in certain restricted cases (discussed below). We are interested here in pronouns and demonstratives. With a few exceptions, inferrables cannot be referred to by pronouns or demonstratives unless they have previously been referred to by a full NP. For the purposes of our model, an NP such as *the front door* should introduce a new entity in the same way as the NP *a house*.

In the algorithm presented here, we will make use of the notion of discourse model in order to simulate pronoun and demonstrative resolution. We do not intend to present a comprehensive model of the discourse. Our simplified model consists only of a list containing representations of the objects that have been referred to in the discourse by NPs. It is similar to Grosz & Sidner’s attentional state, as it is intended to contain representations of entities which are salient to the participants. We will use Webber’s terminology and call these representations *discourse entities* (Webber, 1979). The list is called *S(alience)-list*, as the entities are ordered according to how salient they are in the discourse. The algorithm resolves pronouns by co-indexing them with the highest-ranked compatible entity in the S-list. The list in our model is incrementally updated as the discourse progresses and spans more than one sentence. This means that an entity is available for subsequent anaphoric reference as soon the NP is uttered. The model does therefore not require different mechanisms for inter- and intra-sentential anaphora. The details of the S-list and the resolution process are described in Section 6. We first turn to other linguistic issues.

2.2 Predicate Information

If we say that the referent of an NP is introduced into the discourse model by the NP itself, we can assume that from that point on, the entity in the discourse model is available for subsequent anaphoric reference. We will call this kind of anaphoric reference *individual anaphora*. However, anaphoric reference also occurs with sentential and VP-antecedents (Webber, 1991; Asher, 1993). Following Webber (1991), we will call this *discourse-deictic* reference. In these cases, the matter seems more complex. As can be seen from the following examples (taken from Asher (1993)), anaphors can pick up different kinds of abstract objects such as events, states, concepts, propositions or facts referred to by previous clausal constituents:

(1) **Event:**

John kicked_i Sam on Monday, and it_i hurt. (p.35, 55)

(2) **Concept:**

Somebody [had to take out the garbage,]_i and Bill did it_i. (p.246, 29)

(3) **State:**

John didn't know_i the answer to the problem. This_i lasted until the teacher did the solution on the board. (p.53, 85.b)

(4) **Fact:**

Mary proved [that the defendant was lying about the President's ignorance of the cover-up,]_i This_i shows that the cover-up is much larger than previously thought. (p.245, 28.a)

(5) **Proposition:**

The “liberation” of the village had been bloody. [Some of the Marines had gone crazy and killed some innocent villagers. To coverup the “mistake,” the rest of the squad had torched the village, and the lieutenant called in an air strike.]_i
At first the battalion commander hadn’t believed it_i. (p.49, 82)

Asher states that the type of abstract object is determined by the predicative context of the anaphor. For example, a discourse-deictic anaphor in the subject position of the intransitive verb *hurt* must refer to an event (example 1 above), whereas an anaphor in the object position of the verb *believe* refers to a proposition (example 5 above). In our model, we make use of the predicative context of the anaphor to help distinguish between individual and discourse-deictic anaphors. One can assume that the constituent in the object position of verbs such as *assume* or *believe* refers to an abstract entity and should therefore be co-indexed with a clause. Conversely, the constituent in the object position of the verb *eat* refers to a concrete entity and should therefore be co-indexed with an NP.

It is clear that such a distinction is very simplistic. For example, although the constituent in the object position of *believe* must refer to a proposition, and propositions are generally referred to by whole clauses, this is not always the case. Certain NPs can refer to abstract objects (e.g. *Jane told me [a story]_i. I didn’t believe it_i.) Also, an NP may stand for an abstract object even though it may itself generally refer to a concrete entity. For example, in *I don’t believe Jane*, the NP *Jane* stands for *some/all proposition(s) expressed by Jane*. In spite of these difficulties, we use the*

predicate of the anaphor as one of the features guiding the anaphor classification.

2.3 Referent Coercion

The predicative context of the anaphor is important even when the antecedent constituent has been determined, as the precise referent must still be identified. For example, the clause in (6) can make available an event, concept, proposition or fact as a referent for the anaphor:

- (6) [John [crashed the car]_j]_i.
- (a) This_i annoyed his parents. (event)
 - (b) Jane did that_j, too. (concept)
 - (c) This_i shows how careless he is. (fact)
 - (d) His girlfriend couldn't believe it_i. (proposition)

Instead of assuming that all levels of abstract objects are introduced to the discourse model by the clause that makes them available, it has been suggested that discourse-deictic reference involves *referent coercion* (Dahl & Hellman, 1995) or *ostension* (Webber, 1991). That is, in a process similar to *accomodation* of referents by definite NPs (Lewis, 1979), the anaphor itself creates a new referent in the discourse model.

Webber assumes that referents of discourse-deictic anaphors do not exist in the discourse model unless anaphorically referred to. For each context there are discourse entities that stand proxy for its propositional content. Discourse-deictic anaphora involves a referring function that yields a discourse entity proposition, event, event

type or state from the proxy entity. Passonneau (1991) agrees with Webber, and in addition claims that referents of discourse-deictic anaphors are lost immediately unless referred to again, as seen in the following example:

(7) [I noticed that [Carol insisted on sewing her dresses_k from non-synthetic fabric]_j]._i

That_i's an example of how observant I am.

And they_k always turn out beautifully.

That_j's because she's allergic to synthetics. (p.69)

The demonstrative in the second utterance picks out the referent described in the subordinate clause of the first utterance (*I noticed ...*). The demonstrative in the final utterance cannot be used to refer to a referent in the first utterance (*Carol insisted ...*) because of the intervening discourse-deictic reference (*that_i*). At the time of the fourth utterance the referent of the first utterance is no longer available. For individual anaphoric reference, on the other hand, intervening utterances and anaphoric references pose no such problem. The pronoun *they* is used felicitously in the third utterance to refer to the referent of the NP *her dresses* in the first, in spite of intervening utterances and anaphoric references. Note, however, that chains of discourse-deictic references are possible, as seen in this altered version of Passonneau's example:

(8) [Carol insisted on sewing her dresses from non-synthetic fabric]._i

That_i's because she's allergic to synthetics.

It_i's also because she hates cheap materials.

In (8), the referent of the first clause is available for anaphoric reference both in the second and in the third clause. The continued reference to it assures that it is not lost.

2.4 Choice of NP-form

As we are interested in building a resolution algorithm for pronouns and demonstratives, we now turn to the differences between these NP forms on the one hand, and full NP forms on the other. Gundel et al. (1993), amongst others, note that there is a correlation between different NP forms and the accessibility of their referents. As discussed above, indefinite NPs are often associated with new entities, whereas definite NPs are associated with familiar entities. Gundel et al. make further fine-grained distinctions within the full range of different NP forms. They refer to the degree of accessibility on a scale, which is intended to describe how easy it is for the hearer to identify a referent, that is whether the referent is highly salient and therefore easily retrievable from memory or whether it is less salient and therefore more difficult to retrieve. Their differentiation is more complex than simply *given – new*. The following is an abbreviated version of their accessibility hierarchy and the associated NP forms:

(9) in focus	activated	uniquely identifiable	type identifiable
pronouns	demonstratives	definite full NPs	indefinite NPs

(Adapted from Gundel et al. (1993), p.275)

We will not concern ourselves with the details of their theory, as we are not attempting to build a comprehensive model of the discourse. Instead we will glean some basic distinctions between groups of NP forms from the various theories. If we group together definite and indefinite NPs, as suggested in Section 2.1, we see that on Gundel et al.'s scale, all full NP forms are reserved for entities which are not salient in the discourse model. Gundel et al.'s explanation is that full NPs make enough semantic information explicit to allow correct identification of a non-salient referent. Pronouns are at the opposite end of the accessibility scale. They provide only little information concerning the identity of their referents (in English, number and gender only). They are reserved for the most salient entities in the discourse model, or in Gundel et al.'s terminology, entities that are *in focus*. The demonstratives *this* and *that* are also used for salient entities. The difference between demonstratives and pronouns is that demonstratives indicate that their referent is salient (*activated*), but that it is not the current *most* salient entity (*in focus*).

In the literature, it is generally claimed that discourse-deictic reference, as opposed to individual anaphoric reference, is preferably established with demonstratives rather than pronouns (Webber, 1991; Asher, 1993; Dahl & Hellman, 1995). The contrast in (10) reflects these preferences:

(10) [Jane bought [a new bike]_i]_j.

(a) It_i's great.

(b) That_j's great.

In contexts like this, where the predicate can conceivably be associated with either the referent of a full clause or the NP, the pronoun preferentially picks out an NP antecedent (*a new bike*), whereas the demonstrative picks out the whole clause (*Jane bought a new bike*).

However, contexts where only either an individual or a discourse-deictic interpretation is possible make it clear that both demonstratives and pronouns can be used to make both individual and discourse-deictic reference.

(11) A: I'm going to eat [the last piece of cake]_{*i*}.

B: But John wanted to eat it/that_{*i*}.

(12) A: I wonder whether I should [call him]_{*i*}.

B: I wouldn't do that/it_{*i*} if I were you.

In example 11, the anaphors occur in the object position of the verb *eat*, and must be interpreted as referring to a concrete entity. In example 12, the anaphors occur in the object position of the verb *do* and must thus refer to an event concept.¹

The observation that demonstratives are preferred for discourse-deictic reference is in line with the referent coercion assumption, i.e. the assumption that discourse-deictic anaphoric reference leads to the introduction of a new entity into the discourse model. If one assumes, following Gundel et al., that demonstratives are used for entities that are less salient than those referred to by pronouns, then pronouns are expected to be dispreferred for entities newly created in the discourse model.

2.5 Right Frontier Rule

As with individual anaphora, discourse-deictic anaphora is subject to structural constraints. Webber (1991) notes that only text sections which are on the right frontier of the discourse structure tree are available for discourse-deictic reference, as can be seen by the following discourse (her example 14):

(13) There's two houses you might be interested in.

(a) House A is in Palo Alto. It's got 3 bedrooms and 2 baths, and was built in 1950. It's on a quarter acre, with a lovely garden, and the owner is asking \$425K. But **that**'s all I know about it.

(b) House B is in Portola Valley. It's got 3 bedrooms, 4 baths and a kidney-shaped pool, and was also built in 1950. It's on 4 acres of steep wooded slope, with a view of the mountains. The owner is asking \$600K. I heard all **this** from a real-estate friend of mine.

(c) Is **that** enough information for you to decide which to look at?

(d)*But **that**'s all I know about House A.

The central part of the text is clearly divided into two sections (a and b), each containing the description of a house consisting of more than one clause. At the end of each section a demonstrative is used to refer to what is described by the preceding utterances (*that* for House A; *this* for House B). Finally, in the continuation (c) the demonstrative *that* picks out the referents of the whole preceding discourse, i.e. what is referred to by (13a) and (b) together. The unacceptability of the utterance in the alternative continuation (d) shows that once section (a) is closed off and the

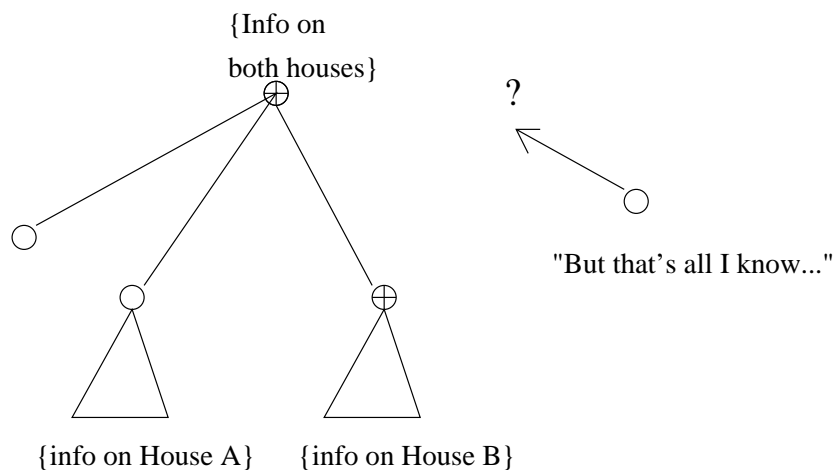


Figure 1: **Discourse Tree Structure (Webber 1991)**

description in section (b) has started, (a) is no longer accessible for reference. Webber represents this discourse with the tree structure shown in Figure 1. The only nodes that a new constituent could attach to are the right-frontier nodes, indicated in the figure by the crossed circles.

Asher's Principle of Availability (Asher, 1993, p.313) has a similar function to the right frontier rule. It states in part that only the current constituent itself and its discourse referents and subconstituents (*subDRSs*) are available as antecedents for abstract object anaphora.² Both Webber's and Asher's findings can be interpreted as reflecting the notion of adjacency. The constituents referred to discourse-deictically must be linearly or hierarchically adjacent to the anaphor. We will make use of this idea in our algorithm, by formulating a concept of adjacency in terms of dialogue acts.

We have so far determined four fundamental differences between anaphoric reference to an NP and discourse-deictic reference. First, abstract objects such as events, states, propositions and facts are not introduced to the discourse-model by virtue

of the constituent that describes them, as in NP reference, but rather by virtue of anaphoric reference. We will call this *referent coercion* (cf. Dahl & Hellman (1995)). The referents of discourse-deictic anaphors are immediately lost again from the discourse model if not referred to again. Secondly, for discourse-deictic reference, the precise referent is determined by the predicate of the anaphor.³ Thirdly, demonstratives are preferentially used for discourse-deictic reference, whereas pronouns are usually used for coreference with NPs. Finally, there are discourse-structural restrictions on discourse-deictic reference in that the antecedent constituents should be linearly or hierarchically adjacent to the anaphor.

Before providing a more detailed description of the algorithm in Section 6, we first describe a preliminary corpus analysis, which was used to test our anaphor classification, coreference method and classification of dialogue acts, and to provide a standard against which the algorithm can be tested.

3 Choice of Corpus

The choice of corpus is a difficult one. All corpora have corpus-specific characteristics which may influence the range of vocabulary and syntactic constructions. The choice should therefore be determined by the specific analysis one wishes to carry out. Our choice to concentrate on spoken rather than written language is guided by previous observations (Eckert, 1998) that written language contains fewer pronominal anaphors and a less diverse range of anaphor types (see Section 4). Furthermore, the purpose of the study is to analyse the effect of grounding on anaphora, and to develop a formal

representation.

Spoken language corpora can roughly be divided into two categories: task-oriented and non-task-oriented. In task-oriented corpora (e.g. TRAINS (Heeman & Allen, 1995), Maptask (Anderson et al., 1991)), the conversational participants are required to perform a particular task, such as construct an object, or describe a route on a map, and are recorded while carrying out the task. The advantage of such corpora is that the common ground between the participants, that is the set of entities familiar to both, is fairly easy to model. The observer can reconstruct whether a particular entity (e.g. *the small screw*) has been previously mentioned, is accessible in the immediate surroundings, or new to the discourse. This feature is particularly valuable when analysing, for example, the appropriate use of the definite article and pronouns. However, such corpora contain a large number of imperative-like constructions, and contain fewer references to non-concrete entities, thus making them unsuitable for our purposes.

Non-task-oriented dialogue corpora are intended to be representative of “natural” and “unconstrained” speech. The Callfriend (LDC, 1996) and Callhome (LDC, 1997) corpora consist of recorded telephone conversations between relatives and friends. These corpora are particularly difficult to analyse as there is a large amount of common ground and shared assumptions between the participants that the observer does not have access to.

For our analysis we chose the Switchboard corpus (LDC, 1993), which is a collection of recorded and transcribed telephone conversations between two people not acquainted with each other. The participants were asked to talk about a given topic,

such as childcare, exercise, or foreign politics. This corpus has some of the advantages of the task-oriented corpora, in that the amount of shared knowledge that is inaccessible to the observer is kept to a minimum. As the dialogues are between strangers, they are easier to follow than those from the Callhome corpus. In addition, the dialogues are not goal-driven and there are many references to both concrete and abstract entities.

4 Anaphora in Dialogues

We now turn to the analysis of anaphora in the corpus. As mentioned in the introductory section, there are anaphors which corefer with NPs, and anaphors which corefer with VPs or clauses. In addition to these two types we identified three other types of pronouns and demonstratives, which do not appear to be corefering with any other linguistic constituent. The correct identification method for anaphors is important, as for the purposes of the algorithm it is necessary to determine which pronouns and demonstratives are anaphoric and therefore resolvable, and which are not. Also, in the case of resolvable anaphors, it is necessary to determine the type of antecedent (NP vs. VP/clause). This section contains a frequency analysis of the different types of pronouns and demonstratives and gives examples of each type from the Switchboard corpus. An empirical analysis of the inter-coder agreement for this classification is presented later in Section 7.

4.1 Individual Anaphors

In the Switchboard corpus dialogues we examined, *individual* anaphors, i.e. anaphors with NP antecedents, constitute only 45.1% of all anaphoric references. This number includes all demonstratives and all instances of *he*, *she*, *it* and *they* with NP antecedents, e.g.

- (14) A: **my parents**_{*i*} didn't really have music in the house . Put it that way .
B: Oh , rea- , Were **they**_{*i*} religious ? (sw4168)

4.2 Reference to Abstract Objects

We classified 22.6% of all anaphors in the corpus as *discourse-deictic*. The referents of discourse-deictic anaphors are abstract objects, such as events, states, event concepts, facts and propositions. Their antecedents are VPs, clauses or sequences of clauses, e.g.

- (15) Now why didn't she [**take him over there with her**]_{*i*}? No, she didn't do **that**_{*i*}. (sw4877)

- (16) A: ...[**we never know what they're thinking**]_{*i*}.

B: **That**_{*i*}'s right. [**I don't trust them**]_{*j*}, maybe I guess **it**_{*j*}'s because of what happened over there with their own people, how they threw them out of power...
(sw3241)

In Example (15) the demonstrative refers to the event concept referent of the preceding VP. In (16), the demonstrative refers to the proposition expressed by the

preceding main clause, and the pronoun *it* refers to the state expressed by the clause *I don't trust them*.

Whilst there have been attempts to classify abstract objects and describe the rules governing anaphoric reference to them (Webber, 1991; Asher, 1993; Dahl & Hellman, 1995), there have been no empirical studies using actual resolution algorithms. However, as described in section 2, there are some important characteristics of abstract object reference which research in theoretical linguistics has mapped out and that we make use of in our algorithm: referent coercion, preference for demonstratives, right frontier rule, and predicate information (see also Eckert & Strube (1999)).

4.3 Vague Anaphors

We classified a further 13.2% of the anaphors as *vague*, in the sense that the pronoun does not have a clearly defined linguistic antecedent. The entities referred to by vague pronouns are similar in nature to the discourse-deictic entities because they are also abstract. However, these pronouns do not refer to the referent of a sentence or VP but to the general discourse topic, as shown in example 17:

(17) B.29 I mean, the baby is like seventeen months and she just screams.

A.30 Uh-huh.

B.31 Well even if she knows that they're fixing to get ready to go over there.

They're not even there yet –

A.32 Uh-huh.

B.33 you know.

A.34 Yeah. **It's** hard. (sw4877)

The pronoun in A.34 is not referring to the specific incidence described by speaker B, but rather to the topic of childcare in general. With these pronouns it is impossible to identify a linguistic string in the context that the pronoun is co-specifying with. An algorithm which relies on linguistic surface form can therefore not resolve them and it is important that they be identified.

In our analysis of the Switchboard dialogues, we observed an interesting contrast. Pronouns appear to be preferred for vague reference, where the referent is not easily identifiable, whereas demonstratives appear to be preferred for clearly defined reference. Note, for example, that in (17) above, if a demonstrative is substituted for the pronoun in A.34, yielding *That's hard*, then it would be interpreted as referring not to the general topic of childcare, but rather to the *specific* incidence described by Speaker B.

4.4 Inferrable-Evoked Pronouns

The remaining 19.1% of anaphors constitute a particular usage of the third person plural pronoun *they*, where it has no explicit antecedent but is associated with a singular NP, e.g.,

(18) A.20 ...in **the Soviet Union**, **they** spent more money on, um, what do you call, um, military power than anything. (sw3241)

In this example, the singular NP *the Soviet Union* has the inferrable *inhabitants/population* associated with it. The highlighted pronoun refers to the inferrable

despite the inferrable itself not having been mentioned explicitly. We call these *Inferrable-Evoked Pronouns (IEP)*. It is usually the case that the NP in question refers to a country, a school, a hospital or some other kind of institution. The pronoun then refers to the authority or the population/members of the institution. Subsets of this type of pronoun have elsewhere been termed *corporate* pronouns (Jaeggli, 1986; Belletti & Rizzi, 1988). Our group of IEP's also includes cases where there is no explicitly mentioned institution, e.g.,

(19) A.19 **They** had an interview with ... The general. Stormin Norman...

A.21 Anyway, at the end of it, **they** rolled all of the U S names of the U S casualties – (sw2403)

The plural pronouns in A.19 and A.20 refer to the television authorities without the institution itself having been mentioned. It seems that certain institutions are salient enough that they require no explicit mention.

IEP's and vague pronouns are the default classes in our algorithm for third person plural pronouns and third person singular neuter pronouns, respectively. They are classified as such by default when the algorithm fails to find a compatible antecedent within a predetermined domain. This is described in detail in Section 5.

4.5 Unmarked Anaphors

We do not mark non-referring pronouns and demonstratives such as expletives, subjects of weather verbs (*quasi-arguments*, (Chomsky, 1981, p.37)) and subjects of raising verbs. Also, we ignore first and second person pronouns, as the correct resolution

of these would require an analysis of deictic shift, which the algorithm is not capable of modelling at this point. More difficult to categorise are the pronouns referred to by Postal & Pullum (1988) as *subcategorised expletives*, which they define as being non-referring pronouns in argument positions, e.g.,

(20) I resent **it** greatly that you didn't call me. (Postal & Pullum, 1988) (ex. 21h)

Idiomatic uses of *it* are also unmarked as in the following:

(21) When **it** comes to trucks, though, I would probably think to go American.

(sw2326)

(22) I haven't prepared any of my lectures, so I'm going to have to wing **it**/***them**.

("improvise") (Postal & Pullum, 1988) (ex. 47c/d)

To identify non-referring pronouns reliably, we use the criterion of possible question formation. In general, *wh*-questions cannot be formed on non-referring pronouns, e.g., **When what comes to trucks?* **What's raining?* **What seems that John snores?*

5 Building Synchronising Units from Dialogue Acts

As mentioned in Section 2, we are assuming that the utterance of an NP can lead to its referent becoming part of the common ground. A question we had left open is determining when this happens. As Byron & Stent (1998) point out, it is difficult to determine the center of attention in multi-party discourse because the participants may not be focussing on the same entity at a given point. Our hypothesis is that the

attentional state of the discourse participants can be determined by making reference to *dialogue acts*. The term *dialogue act* is derived from *speech act* and is intended to bring to mind the communicative function of an utterance in a conversation. We assume that acknowledgments are used by speakers to indicate that common ground is achieved and can therefore indicate which entities have been entered into the joint discourse model. Dialogue acts are also important for a second reason, namely that they can be used as a unit specifying the domain for potential antecedents.

5.1 Dialogue Act Theories

There are many theories of dialogue acts and we here briefly discuss those relevant to our own model. Our common ground assumptions are based on Clark & Schaefer's (1989) theory of contributions (see also Traum's (1994) *Discourse Units* and Nakatani & Traum's (1999) *Common Ground Units*). In Clark & Schaefer's model, each dialogue act is labelled as a *Presentation* or an *Acceptance*. A *Presentation* and an *Acceptance* jointly form a *Contribution*. However, Clark & Schaefer's dialogue act labels are also used for larger units. Their rules are recursive and an *Acceptance* itself can consist of *Contributions*. As shown in Figure 2 (their Figure 4, p.279), a dialogue can contain a subdialogue for clarification purposes. This feature allows discourse structure to be represented. A further important feature of their model is that a single dialogue act may fulfill multiple functions: it can be both an *Acceptance* of a preceding *Presentation* and be a *Presentation* itself.

Carletta et al. (1997) present a more fine grained approach to dialogue acts in

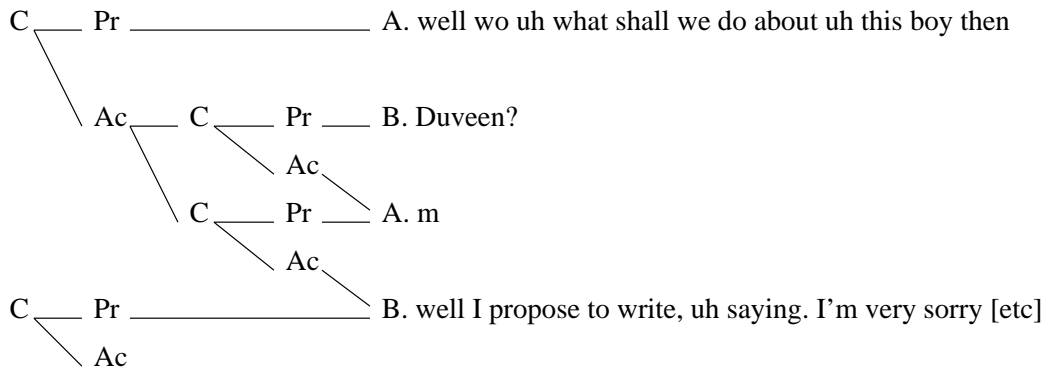


Figure 2: Clark & Schaefer's (1989) Dialogue Structure

their model, which consists of three tiers describing *Moves* (dialogue acts), *Games* (dialogue act sequences) and *Transactions* (subdialogues). Moves are divided into three subtypes – *Initiations*, *Responses* and *Preparations* – and, again, there are numerous subtypes within each of these to capture a variety of different functions.

We wanted our model to fulfill two criteria: (1) it should reflect the achievement of common ground, and (2) it should be simple enough to allow a high degree of inter-coder reliability. To achieve the first, we use pairs of dialogue acts to form *Synchronising Units*, similar but not identical to *Common Ground Units* and *Contributions*. To achieve the second, we simplify Carletta et al.'s model, ignoring the subtypes and using only an *Initiation/Response*-type of distinction. Furthermore, we do not allow for recursive discourse structure, as given in Clark & Schaefer's model.

5.2 Dialogue Acts: Units and Categories in Our Model

We assume that the establishment of common ground is indicated by dialogue acts and affects the operations for adding and removing discourse entities from the representation of the attentional state – in our model the list of salient discourse entities

(S-list). We divide each dialogue into short, clearly defined dialogue acts. As pointed out in Byron & Stent (1998), the determination of utterance boundaries is difficult in spoken language, as annotators must use criteria which do not depend on punctuation. For this reason we define a unit syntactically as:

- each main clause plus any subordinated clauses, or a smaller utterance.

The inclusion of *or a smaller utterance* means that elliptical utterances, which occur frequently in spoken language, can be counted as units. The syntactic constituents serve as an upper boundary for unit definition, but a unit does not need to be syntactically complete.

The labels given to these units are *Initiation* (**I**) and *Acknowledgment* (**A**), based on the top of the hierarchy given in Carletta et al. (1997). **I**'s are dialogue acts that convey semantic content. **A**'s, on the other hand, do not convey semantic content but have the pragmatic function of signalling that the other participant's utterance has been heard or understood. **A**'s are therefore like backchannels. The unit type **A** has an important function and allows us to make use of utterances with no discourse entities, e.g. *Uh-huh; yeah; right*. Whilst Byron & Stent (1998) and Walker (1998) assign no importance to such utterances in their models, in our model, these constitute a specific type of dialogue act which is used to indicate the inclusion of entities into the common ground.

(23) B18 . **I** and it 's just like , everybody likes to blame everything on drugs now ,
I but I wonder , you know ,
I do you get the ,

I oh , that 's kind of side tracked ,

I but , uh , I just remember seeing on the news the other night , they
had the thing about how Catholic schools are doing so much better

A17 . **A** Uh-huh .

(sw3083)

In example 23, we see that Speaker B's turn has been divided into five dialogue acts. His third utterance *do you get the* constitutes a separate unit even though it is less than a full main clause. At the end of B's turn, Speaker A responds with *Uh-huh*. This last dialogue act does not contain any semantic information and is labelled **A**.

Often it is not possible to tease apart **I** and **A**. There are utterances which function as an **A** but also have semantic content, for example answers to wh-questions. This type is labelled as **A/I**. The double label is reminiscent of Clark & Schaefer's model (see Section 5.1), in which a single utterance can fulfill two functions. Expressed in the terms of the dialogue act markup model DAMSL (Allen & Core, 1997; Zollo & Core, 1999), **I**'s are forward-looking in the discourse, **A**'s are backward-looking and **A/I**'s are both forward- and backward-looking. Only forward-looking dialogue acts require a further response or acknowledgment. Table 1 shows a summary of the labelling guidelines from our manual.

5.3 Achieving Common Ground

In order to adequately represent the joint discourse model, we require a further unit which indicates when common ground is achieved. In our model, a single **I** and an

Label	Unit Description	Further Acknowledgment required?
Initiation (I)	Statement Question	Yes (if at turn transition)
Acknowledgment/Initiation (A/I)	Statement following an I Question following an I Answer to a wh-question Answer to a yes/no-question	
Acknowledgment (A)	Vocal signal indicating understanding Word/Phrase indicating understanding	No

Table 1: Guidelines for Labelling Dialogue Acts

A jointly form a *Synchronising Unit (SU)*. Examples of this can be seen in Figure 3. Single **I**'s in longer turns (A.81) constitute **SU**'s by themselves and do not require explicit acknowledgment. The assumption is that by letting the speaker continue, the hearer implicitly acknowledges the utterance. In this sense, **SU**'s differ from Nakatani & Traum's *Common Ground Units* or Traum's *Discourse Units*, which require a response from the other participant to be completed. In our model, it is only in the context of turn-taking that **I**'s and **A**'s are paired up. This is in agreement with Clark & Schaefer's point that "initiation of the relevant next contribution", "acknowledgment" as well as "continued attention" count as evidence of understanding (p. 267).

The **SU**'s have two functions in our model. Firstly, they are used to indicate at which point the S-list is cleaned up – after each **SU**, discourse entities not referred to again are removed from the list. Again, this is a crude simplification but we leave the precise determination of the manner decay of discourse entities for future empirical research. What we wish to supply here is a unit for measuring their duration in the

model. The second point is crucial to our hypothesis that common ground has an influence on attentional state: we assume that at turn transitions only acknowledged **I**'s become part of an **SU**. If at a turn transition one speaker's **I** is not acknowledged by the other participant it cannot be included in an **SU** and its discourse entities are deleted from the S-List.

An example of this latter point is given below in Figure 3. In turn B.84, the entity *our area* is added to the S-List. However, Speaker B is then interrupted by Speaker A. B's **I** is therefore at a turn transition but is not acknowledged. The discourse entity *our area* is then immediately deleted again from the S-List when the subsequent **I** shows that it is not part of the common ground. This means that it is not available as an antecedent for subsequent pronouns. The algorithm correctly predicts that the pronoun in A.85 does not co-specify with *our area*.

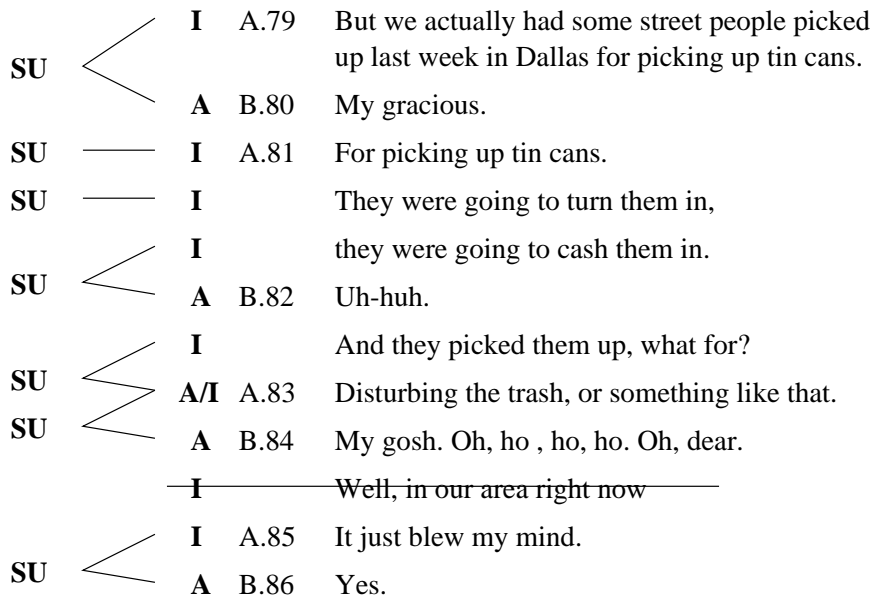


Figure 3: Synchronising Units and Dialogue Acts

5.4 A Note on Incremental Processing

A positive feature of our model (and those such as Traum's) is that, unlike Clark & Schaefer's, it allows the level of dialogue acts to be labelled incrementally. Clark & Schaefer's *Presentations* and *Acceptances* appear not only at the level of dialogue acts but at embedded levels as well, meaning that these labels can only be fully applied to the discourse as a whole.

In our model, labels at the dialogue act level (**I**, **A** and **A/I**) are assigned locally and incrementally, a feature which is compatible with a processing model. At the level of Synchronising Units, labels are also assigned incrementally but retrospective changes can be made. As shown in the examples above, if the content of a particular utterance indicates that the preceding utterance has been ignored, the S-List of the preceding one is deleted and the utterance not included in an **SU**.

The difference between the two levels is due to the fact that the first level represents features of the utterances themselves, whilst the second is an attempt to represent the presuppositions of both speakers. It is unlikely that the presuppositions of all participants are ever identical, so a representation of common ground can only be an approximation. Furthermore, common ground update is generally a feature of more than one utterance, meaning that immediate representation as soon as an utterance is encountered is not feasible.

6 The Algorithm

The method for resolving anaphors in spoken dialogue is based on the incremental algorithm described in Strube (1998). In our method, discourse entities are also added to the S-list (*saliency list*) immediately after they are encountered. The order of the list is based on the information status of the discourse entities, with *hearer-old* discourse

the end of each **SU** all discourse entities which are not realised in this **SU** are removed from the S-list. This means that the size and classification of the dialogue acts determine the set of potential antecedents of an anaphor. A major extension to the Strube (1998) algorithm is the method for the classification of the different types of pronouns and demonstratives presented in Section 4.

In addition to the S-List for individual anaphora, our algorithm also makes use of an A-List, which contains the referents of discourse-deictic anaphors. We also make use of the distinction between demonstratives and pronouns, in particular the preference for demonstratives to be discourse-deictic. Our algorithm consists of two branches, one for pronouns and the other for demonstratives. Both of them call the functions *resolveInd* and *resolveDD*, which resolve individual and discourse-deictic anaphora, respectively. *resolveInd* consists only of a search through the S-list for antecedent matching with respect to gender and number (cf. Strube (1998)).

The function *resolveDD*, on the other hand, consists of a search in the list containing abstract objects, the A-list. It was noted in Section 2, that individual anaphora behave differently from discourse-deictic anaphors, in that the former refer to entities

the S-list, the A-list is cleaned up at the end of each **SU**, meaning that referents which were not referred to again are removed. This reflects Passonneau's (1991) idea the referents of discourse-deictic anaphors are lost immediately after intervening utterances (cf. Section 2).

6.1 Context Ranking – Dialogue Acts and the Right Frontier

Rule

If the A-list is empty (which is usually the case), the algorithm looks through the linguistic context for an appropriate antecedent constituent. The order in which the possibilities are tried out is determined by the *Context Ranking* (examples are given below):

Context Ranking:

- (i) A-list.
- (ii) Within same **I**: Clause to the left of the clause containing the anaphor.
- (iii) Within previous **I**: Rightmost main clause (and subordinated clauses to its right).
- (iv) Within previous **I**'s: Rightmost complete sentence (if previous **I** is incomplete sentence).

If the A-list is empty, the algorithm looks first within the **I** containing the anaphor for the first clause to the left of the anaphor. This is successful in cases such as example (25):

(25) **I** B.104 I hope that, uh, [**they will start picking up on some of these things and, and getting involved**]_i, because **that**_i's the only way that we're going to get out of it. (sw2403)

If there is no clause to the left, as in example (26), it looks to the previous **I** and takes the rightmost main clause. Preceding main and subordinated clauses are ignored.

(26) **I** A.50 because if you tell everybody everything, [**everybody in the world would know because they'd put it on TV**]_i

A B.51 Right.

I A.52 and **that**_i wouldn't do us any good. (sw3241)

In some cases, there is no complete main clause in the preceding **I** alone. The algorithm then looks to all preceding **I**'s until a completed main clause is found. In example (27) (an extract from Figure 3 in the previous section), Speaker A's utterance in A.83 is elliptical but the preceding question in B.82 can be used to form a syntactically complete clause.

(27) **I** B.82 And [**they picked them up, what for?**

A/I A.83 **Disturbing the trash or something like that.**]_i

A B.84 My gosh, Oh, ho, ho, ho. Oh dear.

Well in our area right now,

I A.85 **It**_i just blew my mind. (sw2403)

Because the *Context Ranking* is expressed in terms of dialogue acts, Webber’s *Right-Frontier Rule* (see Section 2) is not violated. Although the text referring to the antecedent is often not *literally* adjacent to the anaphor, it is the adjacent **I** and intervening **A**’s (B.51 in (26) and B.84 in (27)) are invisible for this purpose. Unacknowledged **I**’s, i.e. those not belonging to an **SU** (B.84 in (27)) are also invisible for discourse-deictic reference.

6.2 Anaphor Classification and Resolution

A further point noted in Section 2 is that the predicative context of discourse-deictic anaphors determines what type of abstract object they refer to, i.e. whether they refer to states, events, event concepts, propositions or facts. Our algorithm at present does not have access to the formalised semantic information that would be necessary to make these distinctions explicit but we assume that the predicate of the anaphor creates a referent of the correct type. However, we do use the predicative context of the anaphor to distinguish between individual and abstract anaphors. We define that an anaphor is *I-incompatible* (cannot refer to an individual object) or *A-incompatible*⁴ (cannot refer to an abstract object) if it occurs in one of the corresponding contexts described in Table 2.

An anaphor in the object position of the verb *assume*, for example, is unlikely to have a concrete NP antecedent. This context is therefore listed as being *I-incompatible* in the table. Conversely, the object position of the verb *eat* is unlikely to have an abstract entity such as an event or a proposition as its referent, and the context is

I-Incompatible (*I) Anaphors in the x-position *cannot* refer to individual, concrete entities.

- Equating constructions where a pronominal referent is equated with an abstract object, e.g. *x is making it easy, x is a suggestion*.
- Copula constructions whose adjectives can only be applied to abstract entities, e.g. *x is true, x is correct, x is right*.
- Arguments of propositional attitude verbs, arguments of verbs which *mainly* take S'-complements, e.g. *assume x; say x*.
- Object of *do* (*do x*).
- Anaphoric referent is equated with a "reason", e.g. *x is because I like her*; Anaphor occurs in cleft construction with *how, why*, e.g. *x is why he's late*.

A-Incompatible (*A) Anaphors in the x-position *cannot* refer to abstract entities.

- Equating constructions where a pronominal referent is equated with a concrete individual referent, e.g. *x is a car, x is a nice place to visit*.
- Copula constructions whose adjectives can only be applied to concrete entities, e.g. *x is expensive, x is tasty, x is loud*.
- Arguments of verbs describing physical contact/stimulation, which are generally not used metaphorically, e.g. *break x, smash x, eat x, drink x, smell x, swallow x*.

Table 2: I-Incompatibility and A-Incompatibility

listed as *A-incompatible*.

It is clear that there are problems associated with such lists. One point is that the predicates are in most cases *preferentially* associated with either abstract or individual referents rather than *categorically* (see Section 9 for a discussion of this point). This means that although a predicate may be listed as I-incompatible, it may be that an individual referent is still acceptable in some instances, and vice versa. While the lists do not reflect language use precisely, they do however greatly enhance the performance of the algorithm, as they help avoid a large number of errors.

The majority of predicates are not contained in the lists. Most predicative contexts, e.g. *know x* or *x is good*, allow both concrete and abstract referents in their

argument positions. If an anaphor occurs in such a context, that is, it is neither *I-* nor *A-incompatible*, the classification is determined by the resolution algorithm.

The anaphora resolution algorithm is shown in Tables 3 and 4. If a pronoun (3rd person singular neuter) is encountered (Table 3), the functions *resolveDD* or *resolveInd* are evaluated, depending on whether the pronoun is *I-incompatible* (case 1) or *A-incompatible* (case 2). In the case of success the pronouns are classified as *DDPro* (discourse deictic) or *IPro* (individual), respectively. In the case of failure, the pronouns are classified as *VagPro* (vague). If the pronoun is neither *I-* nor *A-incompatible* (i.e., the predicative context of the pronoun is ambiguous in this respect), the classification is only dependent on the success of the resolution, i.e. on the availability of referents in the S/A-lists. The function *resolveInd* is evaluated first (case 3) because we observed a preference for pronouns to have individual antecedents. If successful, the pronoun is classified as *IPro*, if unsuccessful, the function *resolveDD* attempts to resolve the pronoun (case 4). If this, in turn, is successful, the pronoun is classified as *DDPro*, if it is unsuccessful it is classified as *VagPro*, indicating that the pronoun cannot be resolved using the linguistic context. The procedure is similar in the case of demonstratives (Table 4). The only difference is that (case 3) and (case 4) are reversed to capture the preference for demonstratives to be discourse-deictic (see Section 4).

3rd person masculine or feminine pronouns are resolved directly by a look-up in the S-list as these cannot be discourse-deictic. 3rd person plural pronouns which can be resolved this way are classified as *IPro*, if they cannot be resolved, they are marked as *IEPro* (inferrable-evoked).

1. **case** PRO is I-incompatible
 if *resolveDD*(PRO)
 then classify as *DDPro*
 else classify as *VagPro*
2. **case** PRO is A-incompatible
 if *resolveInd*(PRO)
 then classify as *IPro*
 else classify as *VagPro*
3. **case** PRO is ambiguous
 if *resolveInd*(PRO)
 then classify as *IPro*
 else if *resolveDD*(PRO)
 then classify as *DDPro*
 else classify as *VagPro*

Table 3: Pronoun Resolution Algorithm

1. **case** DEM is I-incompatible
 if *resolveDD*(DEM)
 then classify as *DDDem*
 else classify as *VagDem*
2. **case** DEM is A-incompatible
 if *resolveInd*(DEM)
 then classify as *IDem*
 else classify as *VagDem*
3. **case** DEM is ambiguous
 if *resolveDD*(DEM)
 then classify as *DDDem*
 else if *resolveInd*(DEM)
 then classify as *IDem*
 else classify as *VagDem*

Table 4: Demonstrative Resolution Algorithm

6.3 An Example

The following extract from the corpus (Table 5) is used to exemplify the algorithm. The leftmost column lists the **SU**'s (28- indicates the beginning, -28 the end of the first **SU** in the example), the second column gives the dialogue act labels and the third the speakers and turns. For ease of representation, the S- and A-lists are only given below each **SU** in the state they are at that point in the discourse, and not each time they are updated.

28- -28	I A	B.18 A.19	And [she ₁ ended up going to the [University of Oklahoma] ₂] ₃ . Uh-huh. S: [DE ₁ : she, DE ₂ : Univ. of Oklahoma]
29-29	I	B.20	I can say that ₃ because it ₂ was a big well known school, S: [DE ₂ : it] A: [DE ₃ : that]
30-30	I		it ₂ had a well known education ₄ – S: [DE ₂ : it, DE ₄ : education]

Table 5: Example Analysis

At the end of **SU 28**, the S-list contains the referents of the NPs *she* and *University of Oklahoma*. The demonstrative *that* in turn B.20 is in the object position of the verb *say* and therefore classified as *I-incompatible*. The *Context Ranking* must then determine its referent. There has been no previous discourse-deictic reference so the A-list is empty (or non-existent). There is no clause in the same **I** as the anaphor so it looks to the preceding **I** and gets the referent of the main clause *she ended up going to the University of Oklahoma*. This referent is added to the A-list as *Discourse Entity₃ (DE₃)*.

The first pronoun *it* in B.20 is in an A-incompatible position as the copula construction equates it with a concrete referent (*a big well-known school*). The algorithm searches through the previous S-list for the highest-ranked referent, which in this case is the only referent DE₂.

In **SU 30** there is another pronoun which again is in an A-incompatible context and the S-list must be looked at for an antecedent (DE₂). Through repeated mention this referent is thus kept in the S-list for the entire length of the extract. At the end of **SU 30** no reference has been made to the entity in the A-list (DE₃) so this list is once again empty.

7 Empirical Evaluation

Our data consisted of five randomly selected dialogues from the Switchboard corpus of spoken telephone conversations (LDC, 1993). We empirically evaluated

- the hand annotation of 3 dialogues for dialogue act units, dialogue act labels, classification of pronouns, classification of demonstratives and the co-indexation of anaphors;
- the classification and co-indexation of anaphors in the same 3 dialogues by the algorithm.

Two dialogues were used to train the two annotators (SW2041, SW4877), and three further dialogues for testing hand annotation and algorithm performance (SW2403, SW3117, SW3241).

7.1 Reliability of Hand Annotation

As a measure of inter-coder reliability we used the Kappa-statistic, which was first suggested for linguistic classification tasks by Carletta (1996), and has since been used by others (e.g. Carletta et al. (1997), Passonneau & Litman (1997), Poesio & Vieira (1998)). This statistic measures the percent agreement between annotators but adjusts it by the percent chance agreement for a particular classification task, taking into account the relative frequency of each class. The formula is stated as follows, where PA is the actual agreement between annotators, and PE is the agreement between annotators one would expect by chance:

(28)

$$K = \frac{PA - PE}{1 - PE}$$

A κ of more than .80 is generally assumed to indicate high reliability of the classifications, a κ between .68 and .80 allows tentative conclusions, while a κ lower than .68 shows that the classification is not reliable.

Dialogue Acts. In the first classification task, turns were segmented into dialogue act units. For the purpose of applying the κ statistic we turned the segmentation task into a classification task by using boundaries between dialogue acts as one class, and non-boundaries as the other (see Passonneau & Litman (1997) for a similar practice). Table 6 shows the results. N is the total number of units (boundaries plus non-boundaries), and Z is the total percent agreement, where each unit gets 1 if both annotators agree on its classification and 0 if they do not. The percent agreement (PA) between the annotators was 98.35%, and $\kappa = 0.92$, indicating high reliability of the annotations.

	SW2403	SW3117	SW3241	Σ
Non-Bound.	3372	3332	1717	8421
Bound.	454	452	241	1147
N	1913	1892	979	4784
Z	1877	1866	962	4705
PA	0.9812	0.9863	0.9826	0.9835
PE	0.7908	0.7896	0.7841	0.7890
κ	0.9100	0.9347	0.9200	0.9217

Table 6: Dialogue Act Units

These dialogue act units were then classified into Initiations (I), Acknowledgments (A), Acknowledgment/Initiations (A/I), and no dialogue act (No). For this test we used only those dialogue act units which the annotators agreed about. The PA over labels given to the dialogue act units was 92.6%, $\kappa = 0.87$, again indicating that it is possible to annotate these classes reliably (Table 7).

	SW2403	SW3117	SW3241	Σ
I	230	211	108	549
A	98	120	68	286
A/I	38	41	16	95
No	0	8	8	16
N	183	190	100	473
Z	167	181	90	438
PA	0.9126	0.9526	0.9000	0.9260
PE	0.4774	0.4201	0.4152	0.4273
κ	0.8327	0.9183	0.8290	0.8708

Table 7: Dialogue Act Labels

Individual and Abstract Object Anaphora. For the classification of pronouns (IPro, DDPro, VagPro, IEPro) a PA of 87.5% was measured, $\kappa = 0.81$ (Table 8). For the classification of demonstratives (IDem, DDDem, VagDem) PA was 90.78%, $\kappa = 0.80$ (Table 9).

	SW2403	SW3117	SW3241	Σ
IPro	120	148	5	273
DDPro	33	5	9	47
VagPro	31	20	26	77
IEPro	24	20	86	130
N	104	97	63	264
Z	83	90	58	231
PA	0.7980	0.9278	0.9206	0.8750
PE	0.3935	0.6039	0.5151	0.3571
κ	0.6670	0.8170	0.8363	0.8055

Table 8: Classification of Pronouns

	SW2403	SW3117	SW3241	Σ
IDem	9	19	2	30
DDDem	45	34	28	107
VagDem	5	3	6	14
N	30	28	18	76
Z	27	26	16	69
PA	0.9000	0.9286	0.8888	0.9078
PE	0.5919	0.4866	0.6358	0.5430
κ	0.7550	0.8609	0.6949	0.7985

Table 9: Classification of Demonstratives

Co-Indexation of Anaphora. We used only those anaphors whose classification both annotators agreed upon. The annotators then marked the antecedents and co-indexed them with the anaphors. The results were compared and the annotators agreed upon a reconciled version of the data. Annotator accuracy was then measured against the reconciled version. Table 10 shows that accuracy ranged from 98.4% (Annotator A) to 96.1% (Annotator B) for individual anaphors and from 85.7% to 94.3% for abstract anaphors.

		SW2403	SW3117	SW3241	Σ
Individual	A				
	Agreement	55	69	3	127
	No Agreement	2	0	0	2
	B				
	Agreement	56	65	3	124
	No Agreement	1	4	0	5
Discourse-deictic	A				
	Agreem.	31	15	14	60
	No Agreeem.	7	2	1	10
	B				
	Agreem.	35	16	15	66
	No Agreeem.	3	1	0	4

Table 10: Annotators’ Agreement about Antecedents of Anaphora against Key

7.2 Performance of the Algorithm

We then used the reconciled version of the annotation as the key for the individual and abstract anaphora resolution algorithms. For individual anaphors, Precision was 66.2% and Recall 68.2% (Table 11), for discourse-deictic anaphors Precision was 63.6% and Recall 70% (Table 12). The low value for precision indicates that the classification did not perform very well. Only few of the anaphors resolved incorrectly were classified correctly. One of the most common errors was that a discourse-deictic or vague

anaphor was classified as individual because an individual antecedent was available. A source of errors with respect to the resolution was that we did not allow the domain of the antecedent to exceed one SU. However, exactly this restriction allowed us to resolve many of the discourse-deictic anaphors and also classify a high percentage of *VagPros* and *IEPros* correctly.

	SW2403	SW3117	SW3241	Σ
No. Resolved Correctly	35	52	1	88
No. Resolved Overall	50	77	6	133
No. Resolved in Key	57	69	3	129
Precision	0.7	0.675	0.167	0.662
Recall	0.614	0.754	0.333	0.682

Table 11: Results of the Individual Anaphora Resolution Algorithm

	SW2403	SW3117	SW3241	Σ
No. Resolved Correctly	25	11	13	49
No. Resolved Overall	38	19	20	77
No. Resolved in Key	38	17	15	70
Precision	0.658	0.579	0.65	0.636
Recall	0.658	0.647	0.867	0.7

Table 12: Results of the Discourse-deictic Anaphora Algorithm

8 Comparison to Related Work

Both Webber (1991) and Asher (1993) describe the phenomenon of abstract object anaphora and describe restrictions on the set of potential antecedents. They do not, however, concern themselves with the problem of how to classify a particular pronoun or demonstrative as individual or abstract. Also, as they do not give preferences on the set of potential candidates, their approaches are not intended as attempts to resolve abstract object anaphora.

To our knowledge, only little research has been carried out in the area of anaphora resolution in dialogues. LuperFoy (1992) does not present a corpus study, meaning that statistics about the distribution of individual and abstract object anaphora or about the success rate of her approach are not available. Byron & Stent (1998) present extensions of the centering model (Grosz et al., 1995) for spoken dialogue and identify several problems with the model. However, they also do not present data on the resolution of pronouns in dialogues and do not mention abstract object anaphora. More recently, Zollo & Core (1999) presented their work on the extraction of grounding tags (which correspond to Nakatani & Traum's (1999) Common Ground Units) from dialogue tags. Their work is based on the same idea as ours, that Common Ground Units/ Synchronising Units can be derived from dialogue acts.

9 Conclusions and Future Work

We consider the work presented here to make important contributions to the study of anaphora in two respects. First, we have presented a model of anaphora resolution in spontaneous spoken dialogues. In particular, we have provided a method of structuring dialogues using dialogue acts to define the domain for potential antecedents, thus avoiding the problems that incomplete utterances, repetitions, false starts and utterances with no content words present for methods relying purely on syntactic units. Secondly, we have provided a classification system for the different types of pronouns and demonstratives found in spoken language. This makes it possible to state from the outset which ones are in principle resolvable and which ones do not

have linguistic antecedents. Furthermore, the empirical analysis has drawn attention to the large number of pronouns with non-NP antecedents and with no linguistic antecedents.

For the field of computational linguistics, we hope to have provided a basis for the application of resolution algorithms to spoken language. An important contribution in this respect, is the observation that only two of the pronoun and demonstrative types identified by us are resolvable. Individual anaphors, i.e. those with NP antecedents, have been dealt with by most existing algorithms. We have identified some important criteria which can be used to resolve the second type, i.e. those involving discourse deixis. Our algorithm uses information supplied by the anaphor's predicate as well as the form of the anaphor itself (pronoun vs. demonstrative) to distinguish discourse-deictic from individual reference. For the resolution process of discourse-deictic references, dialogue acts are again used to function as antecedents. We have shown that a model based on these criteria is viable.

We have also identified weak points in the model which could be addressed by future research. As mentioned in Section 6, our use of predicative information does not adequately reflect language use, as it generalises over preferences by making a binary distinction between verbal argument positions requiring individual and abstract object reference. While this allows the algorithm to distinguish many instances of individual and abstract anaphora, the overgeneralisation also results in some mistakes. The errors result primarily for two reasons. The first is that some verbs can be used metaphorically so that a *physical contact* verb such as *swallow*, which we list as A-incompatible can have abstract object anaphors in their argument positions, e.g.,

I told him that [he'd been fired]_i and he swallowed it_i. Secondly, in our anaphor classification, individual anaphors are those co-indexed with NPs, and discourse-deictic anaphors are those co-indexed with VPs and clauses. This is a syntactic distinction. Our distinction between A- and I-incompatible contexts, on the other hand, is semantic, separating abstract from concrete referents. Whilst there is a correlation between NPs and concrete referents on the one hand and between clauses and abstract referents on the other, there are exceptions. Most notably, there are many NPs which refer to abstract entities, and which can therefore function as antecedents for anaphors in so-called A-incompatible verbal contexts, such as the event-referring subject position of *happen*, e.g., *The accident_i . . . It_i happened yesterday.*

To improve this situation, we are currently looking at the possibility of linking the algorithm to a lexical database such as WordNet (see Fellbaum (1998)) to provide semantic information. In WordNet, the NP *accident* (Sense 1), for example, is listed as a hyponym of *event*, thus explaining why it can act as an antecedent for an anaphor we predict to require an event referent:

(29)

accident – (a mishap; especially one causing injury or death)

=> mishap, misadventure, mischance – (an instance of misfortune)

=> misfortune, bad luck – (unnecessary and unforeseen trouble)

=> trouble – (an event causing distress or pain; “what is the trouble?”)

=> happening, occurrence, natural event – (an event that happens)

=> **event** – (something that happens at a given place and time)

An additional problem is that as was pointed out in Section 4, there are different types of abstract objects that discourse-deictic anaphors can refer to. Currently our algorithm does not distinguish between events, states, propositions and facts in the A-List. We assume, following Asher (1993), that the anaphor and its predicate select a referent of the correct type. It is clear, though, that not any clause can function as antecedent for a discourse-deictic anaphor. A clause describing a state, for example, cannot function as an antecedent for an event anaphor, e.g. **[Mary knows French.]_i That_i happens frequently*. We have noted in our corpus that some discourse-deictic anaphors are not immediately adjacent to their antecedents but that such anaphor-antecedent compatibility eliminates potential ambiguity. Providing the algorithm with this kind of information could be useful for selecting the correct antecedent. However, the distinction between events and states involves a complex interaction between lexical information, tense and aspect (cf. Moens & Steedman (1988)), making it difficult to determine simple rules useable in an automated process.

To our knowledge, pronoun resolution algorithms have so far not been applied to the domain of spoken language. Issues such as the number of dialogue acts functioning as the antecedent domain and the characteristics of the entities in the A-List are problems that must be solved empirically. We hope to have provided a solid basis for further work in this area by identifying the specific problems and pointing towards possible solutions.

Acknowledgments. We would like to thank Donna Byron and Amanda Stent for discussing the central issues in this paper and two anonymous reviewers for helpful

comments. We are also grateful for feedback from the participants of Ellen Prince's Discourse Analysis Seminar and the audiences at the Amstelogue '99 workshop and at the Linguistics Research Department, Bell Labs, Lucent Technologies. This work was funded by post-doctoral fellowship awards from IRCS (NSF SBR 8920230).

Notes

¹Although some NPs can function as antecedents to pronouns in the object position of *do* (e.g. *do it/the foxtrot, do drugs*), there is no number and gender compatible antecedent in the preceding clause in example 12.

²This principle also states that a constituent which stands in a discourse relation to the current constituent is available as an antecedent. However, in the simple algorithm we present here we do not deal with discourse relations and so do not make use of this part of the principle.

³We are not claiming that the predicate of an anaphor coreferring with an NP cannot be crucial for disambiguation. However, with NP-anaphoric reference, the predicate does not add entities to the discourse model, but rather it may serve to select one of an already existing group.

⁴The A and I in this terminology should not be confused with the A and I used to refer to Acknowledgements and Initiations – this is a coincidence.

References

- Allen, James F. & Mark Core (1997). *DAMSL: Dialog act markup in several layers*.
Draft of manual, March 1997.
- Anderson, A., M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard,
J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson & R. Weinert (1991).
The HCRC Map Task Corpus. *Language and Speech*, 34:351–366.
- Asher, Nicholas (1993). *Reference to Abstract Objects in Discourse. Series Studies in
Linguistics and Philosophy, Vol. 50*. Dordrecht London:Kluwer Academic.
- Belletti, Adriana & Luigi Rizzi (1988). Psych verbs and theta theory. *Natural Lan-
guage and Linguistic Theory*, 6:291–352.
- Byron, D. & A. Stent (1998). A preliminary model of Centering in dialog. In *Pro-
ceedings of the 17th International Conference on Computational Linguistics and
36th Annual Meeting of the Association for Computational Linguistics, Montréal,
Québec, Canada, 10-14 August 1998.*, pp. 1475–1477.
- Carletta, Jean (1996). Assessing agreement on classification tasks: The Kappa statis-
tic. *Computational Linguistics*, 22(2):249–254.
- Carletta, Jean, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-
Sneddon & Anne Anderson (1997). The reliability of a dialogue structure coding
scheme. *Computational Linguistics*, 23:13–32.

- Chomsky, Noam (1981). *Lectures on Government and Binding*. Dordrecht, Holland:Foris.
- Clark, Herbert H. & Edward F. Schaefer (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.
- Dahl, Östen & Christina Hellman (1995). What happens when we use an anaphor. In *Presentation at the XVth Scandinavian Conference of Linguistics Oslo, Norway*.
- Eckert, Miriam (1998). *Discourse Deixis and Null Anaphora in German.*, (Ph.D. thesis). University of Edinburgh.
- Eckert, Miriam & Michael Strube (1999). Resolving discourse deictic anaphora in dialogues. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, Bergen, Norway, 8-12 June 1999.*, pp. 37–44.
- Fellbaum, Christiane (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.
- Grice, P. (1975). William James lectures on logic and conversation. In D. Davidson & G. Harman (Eds.), *The Logic of Grammar*, pp. 64–75. Encino, Calif.: Dickenson.
- Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein (1995). Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21:203–225.

- Grosz, Barbara J. & Candace L. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204.
- Gundel, Jeanette K., Nancy Hedberg & Ron Zacharski (1993). Cognitive status and the form of referring expressions. *Language*, 69:274–307.
- Heeman, Peter A. & James F. Allen (1995). *The Trains spoken dialogue corpus*. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.
- Heim, Irene R. (1982). *The Semantics of Definite and Indefinite Noun Phrases*, (Ph.D. thesis). University of Massachusetts. PhD thesis published by Graduate Linguistics Student Association, Massachusetts.
- Jaeggli, Osvaldo (1986). Arbitrary plural pronominals. *Natural Language and Linguistic Theory*, 4:43–76.
- Kamp, Hans & Uwe Reyle (1993). *From Discourse to Logic*. Dordrecht:Kluwer Academic Publishers.
- Karttunen, Lauri (1976). Assertion. In J. McCawley (Ed.), *Syntax and Semantics 7*, pp. 363–385. New York: Academic Press.
- Kripke, Saul (1979). Speaker’s reference and semantic reference. In P. French, T. Uehling & H. Wettstein (Eds.), *Contemporary Perspectives in the Philosophy of Language*, pp. 6–27. Minneapolis: University of Minnesota Press.
- LDC (1993). *Switchboard*. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.

- LDC (1996). *CALLFRIEND American English*. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.
- LDC (1997). *CALLHOME American English Speech*. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.
- Lewis, D. (1979). Keeping in a language game. In R. Baeuerle et al. (Ed.), *Semantics from a Different Point of View*. Berlin:Springer Verlag.
- LuperFoy, Susann (1992). The representation of multimodal user interface dialogues using discourse pegs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics; Newark, Delaware, USA, 28 Jun - 2 Jul 1992*, pp. 22–31.
- Moens, Marc & Mark Steedman (1988). Temporal ontology and temporal reference. *Computational Linguistics*, 14:15–60.
- Nakatani, Christine H. & David Traum (1999). A two-level approach to coding dialogue for discourse structure: Activities of the 1998 DRI working group on higher-level structures. In *Proc. of the ACL '99 Workshop Towards Standards and Tools for Discourse Tagging, College Park, Md, June, 1999*, pp. 101–108.
- Passonneau, Rebecca J. (1991). Some facts about centers, indexicals, and demonstratives. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 1991*, pp. 63–70.
- Passonneau, Rebecca J. & Diane Litman (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

- Poesio, Massimo & Renata Vieira (1998). Contributing to discourse. *Computational Linguistics*, 24(2):183–212.
- Postal, Paul & Geoffrey Pullum (1988). Expletive noun phrases in subcategorized positions. *Linguistic Inquiry*, 19:635–670.
- Prince, Ellen (1981). Topicalization, focus-movement, and Yiddish-movement: A pragmatic differentiation. *BLS*, 7:249–64.
- Prince, Ellen (1992). The ZPG letter: subjects, definiteness, and information status. In S. Thompson & W. Mann (Eds.), *Discourse description: Diverse analyses of a fundraising text*. Amsterdam/Philadelphia:John Benjamins.
- Russell, Bertrand (1905). On denoting. *Mind*, 14:479–493.
- Stalnaker, R. (1974). Pragmatic presuppositions. In M. Munitz & P. Unger (Eds.), *Semantics and Philosophy*, pp. 197–213. New York: New York University Press.
- Stalnaker, R. (1979). Assertion. In P. Cole (Ed.), *Syntax and Semantics 9 – Pragmatics*, pp. 315–332. New York: Academic Press.
- Strube, Michael (1998). Never look back: An alternative to centering. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Quebec, Canada, 10-14 August 1998.*, Vol. 2, pp. 1251–1257.

- Traum, David R. (1994). *A Computational Theory of Grounding in Natural Language Conversation*, (Ph.D. thesis). Department of Computer Science, University of Rochester.
- Walker, Marilyn A. (1998). Centering, anaphora resolution, and discourse structure. In Marilyn A. Walker, Aravind K. Joshi & Ellen Prince (Eds.), *Centering Theory in Discourse*. Oxford:Oxford University Press.
- Webber, Bonnie (1991). Structure and ostention in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6:107–135.
- Webber, Bonnie L. (1979). *A formal approach to discourse anaphora*. New York: Garland Pub.
- Zollo, Teresa & Mark Core (1999). Automatically extracting grounding tags from BF tags. In *Proc. of the ACL '99 Workshop Towards Standards and Tools for Discourse Tagging, College Park, Md, June, 1999*, pp. 109–114.

Miriam Eckert,

Institute for Research in Cognitive Science

University of Pennsylvania

3401 Walnut Street, Suite 400A

Philadelphia, PA 19104

USA

`miriame@linc.cis.upenn.edu`

Michael Strube,

European Media Laboratory GmbH

Villa Bosch

Schloss-Wolfsbrunnenweg 33

69118 Heidelberg

GERMANY

`Michael.Strube@eml.villa-bosch.de`