Resolving Discourse Deictic Anaphora in Dialogues

Miriam Eckert & Michael Strube
Institute for Research in Cognitive Science
University of Pennsylvania
3401 Walnut Street, Suite 400A
Philadelphia, PA 19104, USA
{miriame,strube}@linc.cis.upenn.edu

Abstract

Most existing anaphora resolution algorithms are designed to account only for anaphors with NP-antecedents. This paper describes an algorithm for the resolution of discourse deictic anaphors, which constitute a large percentage of anaphors in spoken dialogues. The success of the resolution is dependent on the classification of all pronouns and demonstratives into individual, discourse deictic and vague anaphora. Finally, the empirical results of the application of the algorithm to a corpus of spoken dialogues are presented.

1 Introduction

Most anaphora resolution algorithms are designed to deal with the co-indexing relation between anaphors and NP-antecedents. In the spoken language corpus we examined – the Switchboard corpus of telephone conversations (LDC, 1993) – this type of link only accounts for 45.1% of all anaphoric references. Another 22.6% are anaphors whose referents are not individual, concrete entities but events, facts and propositions, e.g.,

(1) B.7: [We never know what they're thinking]$_i$.
A.8: **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)

Whilst there have been attempts to classify abstract objects and the rules governing anaphoric reference to them (Webber, 1991; Asher, 1993; Dahl and Hellman, 1995), there have been no exhaustive, empirical studies using actual resolution algorithms. These have so far only been applied to written corpora. However, the high frequency of abstract object anaphora in dialogues means that any attempt to resolve anaphors in spoken language cannot succeed without taking this into account.

Summarised below are some issues specific to anaphora resolution in spoken dialogues (see also Byron and Stent (1998) who mention some of these problems in their account of the Centering model (Grosz et al., 1995)).

Center of attention in multi-party discourse. In spontaneous speech it is possible that the participants of a dialogue may not be focussing on the same entity at a given point in the discourse.

Utterances with no discourse entities. E.g., *Uh-huh; yeah; right.* Byron and Stent (1998) and Walker (1998) assign no importance to such utterances in their models. We assume that these also can be used to acknowledge a preceding utterance.

Abandoned or partial utterances. Speakers may interrupt each other or make speech repairs, e.g.,

(2) Uh, our son$_i$ has this kind of, you know, he$_i$'s, well he$_i$ started out going Stephen F Austin (sw3117)

Self-corrected speech cannot be ignored as can be seen by the fact that the entity referred to by the NP *our son* is subsequently referred to by a pronoun and must therefore have entered the discourse model.

Determination of utterance boundaries. Most anaphor resolution algorithms rely on a syntactic definition of utterance which cannot be provided by spoken dialogue as there is no punctuation to mark complete sentences.

These issues are dealt with by our method of segmenting dialogues into dialogue acts with specified discourse functions. In addition, our approach presents a simple classification of individual and abstract object anaphors and uses separate algorithms for each class. We build on the recall rate of state-of-the-art pronoun resolution algorithms but we achieve a far higher precision than would be achieved by applying these to spoken language because the classification of

anaphors prevents the algorithm from co-indexing discourse deictic anaphora with individual antecedents.

Section 2 gives definitions and frequency of occurrence of the different anaphor types. Section 3 describes the segmentation of the dialogues into dialogue acts and the influence of these on the entities in the discourse model. Section 4 presents the method we use for resolving anaphors and the corresponding algorithm. In Section 5, we report on the corpus annotation and the evaluation of the algorithm.

2 Anaphor Types in Dialogues

In the dialogues examined, only 45.1% of the anaphors are **individual anaphors**, i.e., anaphors with NP-antecedents (IPro, IDem), e.g.,

(3) Boeing ought to hire **him**$_i$ and give **him**$_i$ a junkyard$_j$, ...and see if **he**$_i$ could build a Seven Forty-Seven out of **it**$_j$. (sw2102)

22.6% of the anaphors are **discourse deictic**, i.e. co-specify with non-NP constituents such as VPs, sentences, strings of sentences (DDPro, DDDem; cf. Webber (1991)). The phenomenon of discourse deictic anaphora in written texts has been shown to be strongly dependent on discourse structure. As can also be seen in the examples below, anaphoric reference is restricted to elements adjacent to the utterance containing the anaphor, i.e., those on the *right frontier* of the discourse structure tree (Webber, 1991; Asher, 1993):

(4) A.46: [The government don't tell you everything.]$_i$
B.47: I know **it**$_i$
(sw3241)

(5) Now why didn't she [take him over there with her]$_i$? No, she didn't do **that**$_i$.
(sw4877)

The existence of abstract object anaphora shows that aside from individual entities, the discourse model may also contain complex, higher-order entities. One of the differences between individual and discourse deictic anaphora is that whereas a concrete NP antecedent usually only refers to the individual it describes, a sentence may simultaneously denote an eventuality, a concept, a proposition and a fact.

Instead of assuming that all levels of abstract objects are introduced to the discourse model by the sentence that makes them available, it has been suggested that anaphoric discourse deictic reference involves referent *coercion* (Webber, 1991; Asher, 1993; Dahl and Hellman, 1995). This assumption is further justified by the fact that discourse deictic reference, as opposed to individual anaphoric reference, is often established by demonstratives rather than pronouns. In theories relating cognitive status and choice of NP-form (cf. Gundel et al. (1993)), pronouns are only available for the most salient entities, whereas demonstratives can be used to shift the focus of attention to a different entity.

A further 19.1% of anaphors are **Inferrable-Evoked Pronouns** (IEPro) and constitute a particular type of plural pronoun which indirectly co-specifies with a singular antecedent. This group includes *existential*, *generic* and *corporate* 3rd person plural pronouns (Jaeggli, 1986; Belletti and Rizzi, 1988).

(6) I think the *Soviet Union* knows what we have and knows that we're pretty serious and if **they** ever tried to do anything, we would, we would be on the offensive. (sw3241)

In (6), the NP *Soviet Union* can be associated with inferrables such as *the population* or *the government*. These can subsequently be referred to by pronouns without having been explicitly mentioned themselves. In some cases of IEPro's there is no associated NP, as in the following example, where the speaker is referring to the organisers of the Switchboard calls:

(7) this is the first call I've done [...] and, I didn't realize that **they** ha-, were going to reach out to people from [...] all over the country. (sw2041)

13.2% of the anaphors are **vague** (VagPro, VagDem), in the sense that they refer to the general topic of conversation and, as opposed to discourse deictic anaphors, do not have a specific clause as an antecedent, e.g.,

(8) 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.

Non-referring pronouns, or **expletives**, were not marked. These include subjects of weather verbs, those in raising verb constructions or those occurring in sentences with extraposed sentential subjects or objects, e.g.,

(9) **It**'s hard to realize, that there are places that are just so, uh, bare on the shelves as there. (sw2403)

This group also contains the various *subcategorised expletives* (Postal and Pullum, 1988), defined as being non-referring pronouns in argument positions, e.g.,

(10) Uh, they don't need somebody else coming in and saying, you know, okay we're going to be with them and we're going to zap **it** to you. (sw2403)

(11) When **it** comes to trucks, though, I would probably think to go American. (sw2326)

They differ from referring anaphors in that they cannot be questioned (e.g., **When what comes to trucks?*).

3 Synchronising Units

The domain which contains potential antecedents is not given in syntactic terms in spoken dialogue. Hence we define this domain in pragmatic terms. We assume that discourse entities enter the joint discourse model and are available for subsequent reference when common ground between the discourse participants is established. Our model builds on the observation that certain dialogue acts – in particular acknowledgments – signal that common ground is achieved. Our assumptions are based on Clark's (1989) theory of contributions (cf. also Traum (1994)).

Each dialogue is divided into short, clearly defined dialogue acts – Initiations **I** and Acknowledgments **A** – based on the top of the hierarchy given in Carletta et al. (1997). Each sentence and each conjoined clause counts as a separate **I**, even if they are part of the same turn. **A**'s do not convey semantic content but have a pragmatic function (e.g., backchannel). In addition there are utterances which function as an **A** but also have semantic content – these are labelled as **A/I**.

A single **I** is paired with an **A** and they jointly form a *Synchronising Unit (SU)*. In longer turns, each main clause functions as a separate unit along with its subordinate clauses. Single **I**'s 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. It is only in the context of turn-taking that **I**'s and **A**'s are paired up.

Our model is based on the observation that common ground has an influence on attentional state. We assume that only entities in a complete **SU** are entered into the common ground and remain in the S-list for the duration of a further **SU**. If one speaker's **I** is not acknowledged by the other participant it cannot be included in an **SU**. In this case the discourse entities mentioned in the unacknowledged **I** are added to the S-List but are immediately deleted again when the subsequent **I** clearly shows that they are not part of the common ground.

Figure 1 below, taken from the Trains-corpus (speakers *s* and *u*) illustrates that a missing acknowledgment prevents the discourse model from containing discourse entities from the unacknowledged turn.

SU_i	**I**	s:	so there- the five boxcars of oranges <sil> + that are at- + **S-List: [5 boxcars of oranges]**
SU_j	**A/I**	u:	+ at <sil> + at Corning **S-List: [5 boxcars of oranges, Corning]**
	A	s:	um
–	**I**	u:	okay the orange warehouse <sil> um I + have to + **S-List: [Corning, orange warehouse]**
SU_k	**I**	s:	you need + you need to get five <sil> five boxcars of oranges **there** **S-List: [Corning, 5 boxcars of oranges]**
	A	u:	uh
SU_l	**I**		no they're are already waiting for me **there**

(d92a-4.3)

Figure 1: Unacknowledged Turns

Speaker *u*'s second turn is an **I** which is not followed by an **A**. This means that the entity referred to in that utterance (*orange warehouse*) is immediately removed from the joint discourse model. Thus *there* in the final two turns co-specifies with *Corning* and not the most recent *orange warehouse*.

4 How to Resolve Discourse Deictic Anaphora

We now turn to our method of anaphora resolution, which extends the algorithm presented in Strube (1998), in order to be able to account for discourse deictic anaphora as well as individual anaphora.

4.1 Anaphor-antecedent Compatibility

As indicated in Section 2, information provided by the subcategorisation frame of the anaphor's predicate can be used to determine the type of the referent. In the algorithm, we make use of the notion of anaphor-antecedent *Compatibility* to distinguish between discourse deictic and individual reference. Certain predicates (notably verbs of propositional attitude) require one of their arguments to have a referent whose meaning is correlated with sentences, e.g., *is true, assume* (referred to as *SC-bias verbs* in Garnsey et al. (1997) and elsewhere). Pronouns in these positions rarely have concrete individual NP-antecedents and are generally only compatible with discourse deictic referents. Other argument positions are preferentially associated with concrete individuals (e.g., objects of *eat, smell*) (*DO-bias verbs*). A summary of these predicate types is provided in Figure 2, where *I-incompatible*

I-Incompatible (*I)

- 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 false, x is correct, x is right, x isn't right.*
- Arguments of verbs describing propositional attitude which *only* take S'-complements, e.g., *assume.*
- Object of *do.*
- Predicate or anaphoric referent is a "reason", e.g., *x is because I like her, x is why he's late.*

A-Incompatible (*A)

- Equating constructions where a pronominal referent is equated with a concrete individual referent, e.g., *x is a car.*
- 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 cannot be used metaphorically, e.g., *break x, smash x, eat x, drink x, smell x* but NOT **see x*

Figure 2: I-Incompatibility and A-Incompatibility

means *preferentially* associated with abstract objects and *A-incompatible* means *preferentially* associated with individual objects [1]. Anaphors which are argument positions of the first type are classified as discourse deictic (*DDPro; DDDem*), those in argument positions of the second type are classified as individual anaphora (*IPro; IDem*).

It is clear that predicate information alone is not sufficient for this purpose as there is a large group of verbs which allow both individual and discourse deictic referents (e.g., objects of *see, know*) (*EQ-bias verbs*). In these cases the preference is determined by NP-form of the anaphor (pronoun vs. demonstrative).

4.2 Types of Abstract Antecedents

We follow Asher (1993) in assuming that the predicate of a discourse deictic anaphor determines the type of abstract object. An anaphor in the object position of the verb *do*, for example, can only have a VP (event-concept) antecedent (eg *John [sang]. Bill did* **that** *too.*), whereas an anaphor in the subject position of the predicate *is true* requires a full S (proposition) (eg *[John sang].* **That**'*s true.*). This verbal subcategorisation information is used to determine which part of the preceding **I** is required to form the correct referent.

Following Webber and others, we assume that an abstract object is only introduced to the discourse model by the anaphor itself. In addition to the S-List (Strube, 1998), which contains the referents of NPs available for anaphoric reference, our model includes an A-List for abstract objects. This is only filled if discourse deictic pronouns or demonstratives occur and its contents remain only for one **I**, which is necessary for multiple discourse deictic reference to the same entity.

The following *context ranking* describes the order in which the parts of the linguistic context are accessed:

1. A-List (containing abstract objects previously referred to anaphorically).
2. Within same **I**: Clause to the left of the clause containing the anaphor.
3. Within previous **I**: Rightmost main clause (and subordinated clauses to its right).
4. Within previous **I**'s: Rightmost complete sentence (if previous **I** is incomplete sentence).

Figure 3: Context Ranking

4.3 The Algorithm

The algorithm consists of two branches, one for the resolution of pronouns, the other for the resolution of demonstratives. Both of them call the functions *resolveDD* and *resolveInd*, which resolve discourse deictic anaphora and individual anaphora, respectively.

If a pronoun is encountered (Figure 4, below), the functions *resolveDD* or *resolveInd* (described below) are evaluated, depending on whether the pronoun is I-incompatible (1) or A-incompatible (2). In the case of success the pronouns are classified as *DDPro* or *IPro*, respectively. In the case of failure, the pronouns are classified as *VagPro*. If the pronoun is neither I- nor A-incompatible (i.e., the pronoun is ambiguous in this respect), the classification is only dependent on the

[1] These are preferences and not strict rules because some I-Incompatible contexts are compatible with NPs denoting abstract objects, e.g., *The story/It is true.* and NPs which are used to stand elliptically for an event or state, e.g., *His car/It is the reason why he's late.* This shows that predicate compatibility must ultimately be defined in semantic terms and not just rely on syntactic strings (NP vs. S).

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

Figure 4: Pronoun Resolution Algorithm

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

Figure 5: Demonstrative Resolution Algorithm

success of the resolution. The function *resolveInd* is evaluated first (3) because of the observed preference for individual antecedents for pronouns,. If successful, the pronoun is classified as *IPro*, if unsuccessful, the function *resolveDD* attempts to resolve the pronoun (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 (Figure 5, below). The only difference being that the antecedent of a demonstrative is preferentially an abstract object. The order of (3) and (4) is therefore reversed.

We now turn to the function *resolveDD* (Figure 6, below) (assuming that *resolveInd* resolves individual anaphora and returns *true* or *false* depending on its success). In step (1) the function *resolveDD* examines all elements of the *context ranking* (Figure 3) until the function *co-index* succeeds, which evaluates whether the element is of the right type. Then the function *resolveDD* returns *true*. If the pronoun is an argument of *"do"*, the function *co-index* is tried on the VP of the current element of the *context ranking* (2). If successful, the VP-referent is added to the A-List and the function returns *true*. In (3), *co-index* evaluates whether the pronoun and the current element of the *context ranking* are compatible. In the case of a positive result, the element is added to the A-List and *true* is returned. If all elements of the *context ranking* are checked without success, *resolveDD* returns false (4).

```
resolveDD(PRO) :=
1. foreach element of context ranking do
2.    if (PRO is argument of do)
         then if (co-index PRO with VP of element)
              then add VP to A-List; return true
3.       else if (co-index PRO with element)
              then add element to A-List; return true
4. return false.
```

Figure 6: resolveDD

Example 12 illustrates the algorithm:

(12) B.8: I mean, if went and policed, just like you say, every country when they had squabbles,
A.9: Well,
but we've done **it** before,
B.10: Oh,
I know we have.
A.11: and **it** has not been successful.
(sw2403)

When the pronoun *"it"* in A.9 is encountered, the algorithm determines the pronoun to be I-incompatible (Step 1 in Figure 4), as it is the object argument of the verb *do*. The function *resolveDD* is evaluated. The A-List is empty, so the highest ranked element in *the context ranking* is the last complete sentence in B.8. The pronoun is an argument of *"do"*, therefore gets co-indexed with the VP-referent of the sentence in B.8. The VP is added to the A-List, the function returns true and the pronoun is classified as *DDPro* by the algorithm.

When the next pronoun is encountered, the A-List is empty again because of the intervening sentence (**I**) in B.10. The pronoun is neither I- nor A-incompatible, therefore the algorithm evaluates *resolveInd* (step 3). This fails, since there are no individual antecedents available in B.10 and the algorithm evaluates *resolveDD* in the step (4). The first element in the context ranking is the main clause in A.11 which is co-indexed with the pronoun. The clause-referent is added to the A-List, the function returns true and the algorithm classifies the pronoun as *DDPro*. In this case, the classification is correct but not the resolution, since the pronoun should co-specify with the pronoun in A.9.

5 Empirical Evaluation

In order to test the hypotheses made in the previous sections we performed an empirical evaluation on nat-

urally occurring dialogues. First, the corpus was annotated for all relevant features, i.e., division of turns into dialogue act units, classification of dialogue acts (**I**, **A**), marking of noun phrases, classification of the various types of anaphors introduced in Section 2, and annotating coreference between anaphors and individual/abstract discourse entities. The last step provided the key for the test of the algorithm described in Section 4.3.

5.1 Annotation

Our data consisted of five randomly selected dialogues from the Switchboard corpus of spoken telephone conversations (LDC, 1993). Two dialogues (SW2041, SW4877) were used to train the two annotators (the authors), and three further dialogues for testing (SW2403, SW3117, SW3241). The training dialogues were used for improving the annotation manual and for clarifying the annotation in borderline cases.

After each step the annotations were compared using the κ statistic as reliability measure for all classification tasks (Carletta, 1996). A κ of $0.68 < \kappa < 0.80$ allows tentative conclusions while $\kappa > 0.80$ indicates reliability between the annotators. In the following tables, the rows on above the horizontal line show how often a particular class was actually marked as such by *both* annotators. In the rows below the line, N shows the total number of markables, while Z gives the number of agreements between the annotations. PA is percent agreement between the annotators, PE expected agreement by chance. Finally, κ is computed by the formula $PA - PE/1 - PE$.

Dialogue Acts. First, turns were segmented into dialogue act units. 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 and Litman (1997) for a similar practice). In Table 1, *Non-Bound.* and *Bound.* give the number of non-boundaries and boundaries actually marked by the annotators, N is the total number of possible boundary sites, while Z gives the number of agreements between 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 1: Dialogue Act Units

Table 2 shows the results of the comparison between the annotations with respect to the classification of the dialogue act units into Initiations (I), Acknowledgements (A), Acknowledgement/Initiations (A/I), and no dialogue act (No). For this test we used only these dialogue act units which the annotators agreed about. PA was 92.6%, $\kappa = 0.87$ again indicating that it is possible to annotate these classes reliably.

	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 2: Dialogue Act Labels

Individual and Abstract Object Anaphora. Table 3[2] shows the reliability scores for the classification of pronouns in the classes IPro, DDPro, VagPro, and IEProclassification of demonstratives in the classes IDem, DDDem, and VagDem. The κ-values are around .8, indicating that annotators were able to classify the pronouns reliably.

	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 3: 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 4: Classification of Demonstratives

Co-Indexation of Abstract Object Anaphora. The abstract object anaphora were manually co-indexed

[2]No. for each class is the actual no. marked by *both* annotators. N is the total number of markables, Z is total number of agreements between annotators, PE is the expected agreement by chance.

with their antecedents. For this task we cannot provide reliability scores using κ because it is not a classification task. It is much more difficult than the previous ones, as the problem consists of identifying the correct beginning and end of the string which co-specifies with the anaphor. We used only the abstract 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. Accuracy ranged from 85,7% (Annotator A) to 94,3% (Annotator B).

	SW2403	SW3117	SW3241	Σ
A				
Agreem.	31	15	14	60
No Agreem.	7	2	1	10
B				
Agreem.	35	16	15	66
No Agreem.	3	1	0	4

Table 5: Agreement about Antecedents of Discourse Deictic Anaphora against Key

5.2 Evaluation of the Algorithm

We used the reconciled version of the annotation as key for the abstract anaphora resolution algorithm. Table 6 shows the results of the evaluation. Precision is 63.6% and Recall 70%.

	SW2403	SW3117	SW3241	Σ
Res. Corr.	25	11	13	49
Res. Overall	38	19	20	77
Res. Key	38	17	15	70
Precision	0.658	0.579	0.65	0.636
Recall	0.658	0.647	0.867	0.7

Table 6: Results of the Discourse Deictic Anaphora Algorithm

The low value for precision indicates that the classification did not perform very well. Of the 28 anaphors resolved incorrectly, only 11 were classified correctly. One of the most common errors in classification was, that an anaphor annotated as vague (*VagPro, VagDem*) was classified by the algorithm as discourse deictic (*DDPro, DDDem*). Classification is dependent on resolution, so since the context almost always provides an antecedent for a discourse deictic anaphor, it is possible to classify and resolve a vague anaphor incorrectly, as in Example 13:

(13) A: [I don't know]$_i$, I think **it**$_i$ really depends a lot on the child.
(sw3117)

6 Comparison to Related Work

Both Webber (1991) and Asher (1993) describe the phenomenon of abstract object anaphora and present restrictions on the set of potential antecedents. They do not, however, concern themselves with the problem of how to classify a certain 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.

Concerning anaphora resolution in dialogues, only little research has been carried out in this area to our knowledge. 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 and Stent (1998) present extensions of the centering model (Grosz et al., 1995) for spoken dialogue and identify several problems with the model. We have chosen Strube's (1998) model for the resolution of individual anaphora as basis because it avoids the problems encountered by Byron & Stent, who also do not present data on the resolution of pronouns in dialogues and do not mention abstract object anaphora.

Dagan and Itai (1991) describe a corpus-based approach to the resolution of pronouns, which is evaluated for the neuter pronoun *"it"*. Again, abstract object anaphora are not mentioned.

7 Conclusions and Future Work

In this paper we presented a method for resolving abstract object anaphora in spoken language. We consider our approach to be a first step towards the unconstrained resolution of anaphora in dialogue.

The results of our method show that the recall is fairly high while the precision is relatively low. This indicates that the anaphor classification requires improvement, in particular the notion of *Compatibility*. Lists of verb biases for sentential and NP complements, as described in psycholinguistic studies (e.g. Garnsey et al. (1997)), could be used to classify verbs. Currently exisiting lists only account for a small number of verbs but there may be the possibility of adding statistical information from large corpora of spoken dialogue.

Furthermore, the algorithm currently ignores abstract NPs (e.g., *story, exercising*) when looking for antecedents for anaphors with *I-incompatible* predicates. We are considering determining the feature *abstract* for all NPs in order to identify those which can act as antecedents in such contexts.

Information such as this could be used by the algorithm to prevent the anaphor classification from being dependent on anaphor resolution.

Acknowledgments. We would like to thank Donna Byron and Amanda Stent for discussing the central issues contained in this paper. We are grateful to audiences at AT&T Labs-Research, the University of Delaware, IBM Research and the participants of Ellen Prince's Discourse Analysis Seminar for the critical feedback they provided. Thanks also to Jonathan DeCristofaro and Kathleen F. McCoy who discussed the empirical issues. Both authors are funded by postdoctoral fellowship awards from IRCS.

References

Nicholas Asher. 1993. *Reference to Abstract Objects.* Kluwer, Dordrecht.

Adriana Belletti and Luigi Rizzi. 1988. Psych verbs and theta theory. *Natural Language and Linguistic Theory*, 6:291–352.

Donna Byron and Amanda Stent. 1998. A preliminary model of centering in dialog. In *Proceedings 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, pages 1475–1477.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Ido Dagan and Alon Itai. 1991. A statistical filter for resolving pronoun references. In Y.A. Feldman and A. Bruckstein, editors, *Artificial Intelligence and Computer Vision*, pages 125–135. Elsevier, Amsterdam.

Östen Dahl and Christina Hellman. 1995. What happens when we use an anaphor. In *Presentation at the XVth Scandinavian Conference of Linguistics Oslo, Norway.*

Susan Garnsey, Neal Pearlmutter, Elizabeth Myers, and Melanie Lotocky. 1997. Contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37:58–93.

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

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.

Osvaldo Jaeggli. 1986. Arbitrary plural pronominals. *Natural Language and Linguistic Theory*, 4:43–76.

LDC. 1993. Switchboard. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.

Susann LuperFoy. 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, Del., 28 June – 2 July 1992, pages 22–31.

Rebecca Passonneau and Diane Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Paul Postal and Geoffrey Pullum. 1988. Expletive noun phrases in subcategorized positions. *Linguistic Inquiry*, 19:635–670.

Michael Strube. 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, Québec, Canada, 10–14 August 1998, pages 1251–1257.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation.* Ph.D. thesis, Department of Computer Science, University of Rochester.

Marilyn A. Walker. 1998. Centering, anaphora resolution, and discourse structure. In M.A. Walker, A.K. Joshi, and E.F. Prince, editors, *Centering Theory in Discourse*, pages 401–435. Oxford University Press, Oxford, U.K.

Bonnie L. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.