Paper presentend at Corpus Linguistics 2005, University of Birmingham, England

Annotating (Anaphoric) Ambiguity

Massimo Poesio and Ron Artstein
University of Essex
Language and Computation Group / Department of Computer Science
United Kingdom
{*poesio*|*artstein*}*@essex.ac.uk*

Abstract

We report the results of a preliminary study attempting to identify ambiguous expressions in spoken language dialogues. In this study we developed methods for marking explicit ambiguity, and generalized previous proposals by Passonneau concerning a distance metric for anaphora to be used with the α coefficient to allow for ambiguous annotations.

1 INTRODUCTION

Although it is well-known that natural language expressions can be ambiguous, whether deliberately, as in poetry (Su, 1994) or humour (Raskin, 1985), or unintentionally, few attempts have been made at systematically studying the occurrence of ambiguous expressions in language. Yet, such a study is important both from a linguistic point of view and from an annotation technology point of view: ambiguous expressions may well result in disagreement among coders, and some decision has to be made concerning how to annotate these cases. Consider the dialogue excerpt in (1):[1] it's not clear to us (nor was to our annotators, as we'll see below) whether the demonstrative *that* in utterance unit 18.8 refers to the 'bad wheel' or 'the boxcar'; as a result, annotators' judgments may disagree – but this doesn't mean that the annotation scheme is faulty; only that what is being said is genuinely ambiguous.

[1] This example, like most of those in the rest of the paper, is taken from the first edition of the TRAINS corpus collected at the University of Rochester (Gross *et al.*, 1993). The dialogues are available at `ftp://ftp.cs.rochester.edu/pub/papers/ai/92.tn1.trains_91_dialogues.txt`.

(1) 18.1 S: ....
18.6 it turns out that the boxcar at Elmira
18.7 has a bad wheel
18.8 and they're .. gonna start fixing **that** at midnight
18.9 but it won't be ready until 8
19.1 M: oh what a pain in the butt

However, whereas much attention has been paid in work on discourse to the issue of how to deal with disagreement problems resulting from the subjectivity of the coding schemes, we are not aware of much work addressing the issues arising from ambiguous expressions. In all annotation studies we are aware of,[2] the fact that an expression may not have a unique interpretation in the context of its occurrence is viewed as a problem with the annotation scheme, to be fixed by, e.g., developing suitably underspecified representations, as done particularly in work on wordsense annotation (Buitelaar, 1998; Palmer *et al.*, 2005), but also on dialogue act tagging. Unfortunately, the underspecification solution only genuinely applies to cases of polysemy, not homonymy (Poesio, 1996), and anaphoric ambiguity is not a case of polysemy, as shown by the previous example.

Although we will concentrate here on anaphoric ambiguity, this problem is encountered with all types of annotation; the view that all types of disagreement indicate a problem with the annotation scheme–i.e., that somehow the problem would disappear if only we could find the right annotation scheme, or concentrate on the 'right' types of linguistic judgments–is, in our opinion, misguided. A better approach is to find when annotators disagree because of intrinsic problems with the text, or, even better, to develop methods to identify genuinely ambiguous expressions–the ultimate goal of this work.

In the paper we first discuss the methodology we used in an anaphoric annotation experiment to allow annotators to mark expressions as ambiguous. We then analyze the results in a qualitative way, before considering the problem of measuring agreement in a scheme allowing for ambiguity. Finally, we discuss the implications of this work.

[2] The one exception is Rosenberg and Binkowski (2004).

2 AN EXPERIMENT IN (AMBIGUOUS) ANAPHORIC ANNOTATION

2.1 Annotating Anaphora

As said above, the focus of our research are anaphoric expressions, but at this stage we are not yet proposing a new scheme for annotating anaphora. The coding manual used in this experiment is based on the approach to anaphoric annotation developed in MATE (Poesio *et al.*, 1999) and GNOME (Poesio, 2004), simplifying task and instructions (the primary simplification being that we did not annotate bridging references at this stage), and adding instructions for annotating ambiguous anaphora and a simple way for marking discourse deixis.

The task of 'anaphoric annotation' discussed here is related to, although different from, the task of annotating 'coreference' in the sense of the so-called MUCSS scheme developed for the MUC-7 initiative (Hirschman, 1998). This scheme, while often criticized, is widely used, and has been the basis of coreference annotation for the ACE initiative in the past two years; it suffers however from a number of problems (van Deemter and Kibble, 2000), chief among which is the fact that the one semantic relation captured by the scheme, `ident`, conflates COREFERENCE proper with a number of semantically distinct relations, such as the more general IDENTITY ANAPHORA (for non-referring expressions), BOUND ANAPHORA, and even PREDICATION. (Space prevents a fuller discussion and exemplification of these relations here.)

The goal of the MATE and GNOME schemes (as well of other schemes developed by Passonneau (1997) and Byron (2003)) was to devise instructions appropriate for the creation of resources suitable for the theoretical study of anaphora from a linguistic and psychological perspective, and, from a computational perspective, for the evaluation of anaphora resolution and referring expressions generation. The goal of these schemes is to annotate the DISCOURSE MODEL resulting from the interpretation of a text, in the sense of (Webber, 1979) and of dynamic theories of anaphora such as Discourse Representation Theory (DRT) (Heim, 1982; Kamp and Reyle, 1993). In order to do this, annotators must first of all identify what we call TERMS: the noun phrases that either introduce new discourse entities (DISCOURSE-NEW (Prince, 1992)) or are mentions of previously introduced ones (DISCOURSE-OLD), ignoring noun phrases that are used predicatively.[3] Secondly, annotators have to specify which discourse entities have the same interpretation. Given that the characterization of such discourse models is usually considered part

[3] Our 'terms' correspond to 'referring' noun phrases of functional linguistics (Gundel *et al.*, 1993) and NLG (Dale, 1992); we'll however avoid using the term 'referring' to avoid confusions.

of the area of the semantics of anaphora, and that the relations to be annotated include relations other than Sidner's (1979) COSPECIFICATION, we use the term ANNOTATION OF ANAPHORA for this task (Poesio, 2004), but the reader should keep in mind that we are not only concerned with nominal expressions which are lexically anaphoric.

2.2 Taking Ambiguity into Account

Our theoretical framework for discussing ambiguity, underspecification and related notions is derived from Pinkal (1995) as modified by Poesio (To appear). The most important distinction for the present purposes is that between POLYSEMY and HOMONYMY. Polysemy is the case of ambiguity in which the distinct meanings are somehow related: a typical example is the ambiguity of *mouth* between a sense indicating "the opening through which food is taken in and vocalizations emerge" and "the point where a stream issues into a larger body of water" (both glosses from WordNet 2.0). Polysemy is especially common for wordsenses, particularly of verbs, and is commonly handled by introducing underspecified tags covering several interpretations (Buitelaar, 1998; Palmer *et al.*, 2005). Homonymy, by contrast, is the case of ambiguity for which no common interpretation exists: the classic *bank* is a typical example. The lack of a common interpretation makes the 'underspecified' approach theoretically inappropriate for homonymy cases, which is not a big problem for wordsenses as generally context helps disambiguate homonym words; things are different for anaphora however, as shown below.

In earlier analyses of the TRAINS-91 corpus (Poesio *et al.*, To appear) we identified two types of systematically ambiguous anaphoric expressions in the dialogues of the corpus, which we aimed to study more systematically via annotation. The first class are examples which we called MEREOLOGICAL cases, such as those in (1): anaphoric expressions referring to one of two objects which have been joined together. These expressions are fairly clear cases of homonymy,[4] in the sense that the boxcar and the wheel are clearly distinct objects which we would not want to be part of the same anaphoric chain. The second class of systematically ambiguous expressions are references to plans such as the two uses of demonstrative *that* in utterance units 4.2 and 4.3 of the following transcript fragment, which could refer either to most recently introduced action along the 'right frontier' (picking up the tanker) or to the entire plan proposed in 1.4–3.1.

[4] Pinkal (1995) introduces the terms H-AMBIGUITY and P-AMBIGUITY to refer to the types of ambiguity of which homonymy and polysemy, respectively, are the instantiations for lexical semantics. Forcing the terminology somewhar, we will just use the terms homonymy and polysemy to refer to h-type and p-type ambiguity also for non-lexical ambiguity.

```
(2)  1.4 M: first thing I'd like you to do
     1.5    is send engine E2 off with a boxcar to Corning
            to pick up oranges
     1.6    uh as soon as possible
     2.1 S: okay
     3.1 M: and while it's there it should pick up the tanker
     4.1 S: okay
     4.2    and that can get
     4.3    we can get that done by three
```

The situation with these examples is less clear, but provisionally at least we assume they are cases of homonymy, as well, because the two actions are distinct.

Our approach to annotating both types of ambiguous anaphoric expressions was to ask subjects to mark multiple antecedents, instead of a single underspecified interpretation. A difficulty when trying to do this is the fact that not all ambiguities are detected, at least not immediately. This observation is often found in psycholinguistic experiments, in which the existence of alternative interpretations of a certain expression can only be detected by the fact that different groups of subjects assigned distinct interpretations to it (for an example of implicit ambiguity revealed by analyzing subjects' responses, see (Kurtzman and MacDonald, 1993)). In previous work (Poesio, 1996) we introduced the terms EXPLICIT AMBIGUITY to refer to ambiguity immediately perceived by the subject, and IMPLICIT AMBIGUITY to refer to ambiguity which is only revealed by discrepancies in interpretation. Clearly we can only expect annotators to mark cases in which they detect the ambiguity, i.e., cases of explicit ambiguity.

2.3 The Experimental Setup

Materials. The TRAINS 91 corpus consists of transcripts of dialogues between two humans. One of the humans plays the 'manager' of a railway company, with aim to develop a plan to achieve a transportation goal (delivering a certain amount of goods at a given town by a given deadline). The other participant in the dialogue plays a 'system,' and her role is to help managers develop this plan and provide them with the required information. The text annotated in the experiment was dialogue 3.2 from the TRAINS 91 corpus. Subjects were trained on dialogue 3.1.

Tools. The subjects performed their annotations on Viglen Genie workstations with LG Flatron monitors running Windows XP, using the MMAX 2 annotation tool (Müller and Strube, 2003).[5]

[5] Available from `http://mmax.eml-research.de/`

Subjects. Eighteen paid subjects participated in the experiment, all students at the University of Essex, mostly undergraduates from the Departments of Psychology and Language and Linguistics, and were paid £30 for their participation.

Procedure. The subjects performed the experiment together in one lab, each working on a separate computer, displaying both the text to annotate and a map of the 'TRAINS world'. The experiment was run in two sessions, each consisting of two hour-long parts separated by a 30 minute break. The first part of the first session was devoted to training: subjects were given the annotation manual and taught how to use the software, and then annotated the training text together. After the break, the subjects annotated the first half of dialogue 3.2 (up to utterance 19.6). The second session took place five days later. In the first part we quickly pointed out some problems in the first session (for instance reminding the subjects to be careful during the annotation), and then immediately the subjects annotated the second half of the dialogue, and wrote up a summary. The second part of the second session was used for a separate experiment with a different dialogue and a slightly different annotation scheme.

2.4 Annotation Instructions

The MMAX 2 tool we are using for these experiments allows for multiple types of markables; for this experiment, markables at the phrase, utterance, and turn levels were defined. All noun phrases except temporal ones were treated as phrase markables (Poesio, 2004). Subjects were instructed to go through the phrase markables in order (using MMAX 2's markable browser) and assign each markable to one of four classes: `phrase` if it referred to an object which was mentioned earlier in the dialogue; `segment` if it referred to a plan, event, action, or fact discussed earlier in the dialogue; `place` if it was one of the five railway stations in the 'TRAINS world' (Avon, Bath, Corning, Dansville, and Elmira), and it was explicitly mentioned by name; or `none` if the markable did not fit any of the above criteria, for instance if it referred to a novel object or was not a referential noun phrase.[6] For markables designated as `phrase` or `segment`, subjects were instructed to create a `pointer` to the antecedent, a markable at the phrase or turn level. (See below.) In case an expression was considered ambiguous, subjects were instructed to create more than one pointer. Markables which were not classified, or which were marked `phrase` or `segment` but for which no antecedent was specified, were

[6] We included the value `place` in order to avoid having our subjects mark pointers from explicit place names. These occur frequently in the dialogue–49 of the 151 markables–but are rather uninteresting as far as anaphora goes.

considered data errors; data errors occurred in 3 out of the 151 markables in the dialogue, and these items were excluded from the analysis.

We chose to mark antecedents using MMAX 2's pointers, rather than its sets, because pointers allow us to annotate ambiguity: an ambiguous phrase can point to two antecedents without making them part of the same anaphoric chain. In addition, MMAX 2 makes it possible to restrict pointers to a particular level. In our scheme, markables marked as `phrase` could only point to phrase-level antecedents while markables marked as `segment` could only point to turn-level antecedents, thus simplifying the annotation.

As in previous studies (Eckert and Strube, 2001; Byron, 2003), we only allowed a constrained form of reference to discourse segments: our subjects could only indicate turn-level markables as antecedents. This resulted in rather coarse-grained markings, especially when a single turn was long and included discussion of a number of topics. A more complicated annotation scheme allowing a more fine-grained marking of reference to discourse segments is being tested in a follow-up experiment. The full annotation manual is available upon request.

3 AMBIGUITY IN THE DATA

Our results so far can be divided in two parts: an analysis of the type of ambiguity found in our data (in this section) and results concerning the measurement of agreement on ambiguous data (next section).

3.1 The frequency of ambiguous expressions

The results of the experiment are summarized in Table 1. There was perfect agreement among annotators on 65 / 148 markables (43.9%) and near perfect agreement (no more than 2 disagreeing coders) for another 18 markables (12.2%)—in total, there were no real disagreements on 56.1% of markables. The remaining 63 markables[7] (42.6%) were marked as at least implicitly ambiguous, in the sense that there were at least two antecedents chosen by more than two coders each. Of these 63 markables, 23 (15.5% of the total number of markables) were marked as explicitly ambiguous by at least one annotator. In the first half of the test dialogue, 15 markables out of 72 (20.8%) were marked as explicitly ambiguous, for a total of 55 explicit ambiguity markings (45 phrase references, 10 segment references); in the second, 8/76, 10.5%.

[7] See footnote c) in Table 1.

	First Half	Second half	Total
Number of markables	72	76	148
Perfect agreement	27 (37.5%)	38 (50.0%)	65 (43.9%)
Almost perfect[a]	10 (13.9%)	8 (10.5%)	18 (12.2%)
Ambiguous (total)[b]	35 (48.6%)	28[c] (36.8%)	63 (42.6%)
Explicit ambiguity[d]	15 (20.8%)	8 (10.5%)	23 (15.5%)
Anaphora / DNew[e]	8 (11.1%)	19 (25.0%)	27 (18.2%)

[a] items for which 16 or 17 subjects gave identical judgments

[b] items for which at least two labels were chosen by at least two subjects each

[c] two additional items were assigned a single label by 14 or 15 subjects, and distinct labels by each of the remaining subjects

[d] items which at least one annotator marked as explicitly ambiguous

[e] items ambiguous between a discourse-old and a discourse-new interpretation

Table 1: Ambiguity in the data

3.2 Types of ambiguity

The difference between annotation of (identity!) anaphoric relations and other semantic annotation tasks such as dialogue act or wordsense annotation is that apart from the occasional example of carelessness, such as marking *Elmira* as antecedent for *the boxcar at Elmira*,[8] all other cases of disagreement reflect a genuine ambiguity, as opposed to differences in the application of subjective categories.[9]

The relation between explicit implicit ambiguity is clearly illustrated with reference to the part of the dialogue in (2), repeated in (3).

```
(3)   1.4 M: first thing _I’d_ like you to do
      1.5    is send engine E2 off with a boxcar to Corning
             to pick up oranges
      1.6    uh as soon as possible
      2.1 S: okay [6 sec]
      3.1 M: and while it’s there it should pick up the tanker
```

The two *it* pronouns in utterance unit 3.1 are examples of the type of ambiguity already seen in (1). All of our subjects considered the first pronoun a ‘phrase’

[8] According to our (subjective) calculations, at least one annotator made one obvious mistake of this type for 20 items out of 72 in the first half of the dialogue–for a total of 35 careless or mistaken judgment out of 1296 total judgments, or 2.7%.

[9] Things are different for associative anaphora, see (Poesio and Vieira, 1998).

reference. 9 coders marked the pronoun as explicitly ambiguous between engine E2 and the boxcar; 6 marked it as unambiguous and referring to engine E2; and 3 as unambiguous and referring to the boxcar.

The results for discourse deixis were more complex to discuss, as our annotators clearly had more trouble with this type of references. There was no case of perfect agreement on discourse deixis, but we did find several cases of near perfect agreement. We found a much greater percentage of such cases annotated as explicitly or implicitly ambiguous, but the pattern for cases of ambiguous discourse deixis such as those in (2) was similar to that for the 'mereology' cases: for example, the first *that* in (2) (utterance 4.2) was marked by six coders as referring to the action introduced in 1.4–1.6, three coders as referring to the action in 3.1, and two coders as ambiguous between the two (or possibly as referring to the sum, see below).

Interestingly, the most common example of ambiguity found in the annotation was not one of the cases we had developed methods for marking explicitly: this was the ambiguity between a discourse-new and discourse-old interpretation of indefinites referring to stuff. Although the first mention of the *oranges* in (3) was marked as discourse-new by all of our annotators, with all the subsequent references we found a disagreement between annotators who marked the mention as referring to the same oranges, or to new entities of the same type.

Finally, we found that several coders had problems distinguishing between ambiguity and plurality; in many cases of plural anaphora referring to two or more objects introduced in the dialogue (say, an engine and a boxcar) , these coders used two pointers to mark the two antecedents.

Preliminary conclusions we can draw from the discussion in this section are the need (i) to clarify to coders this last distinction, (ii) for methods for marking the ambiguity between an anaphoric and a non-anaphoric interpretation, and (iii) for methods for identifying ambiguous cases considering not only the cases of *explicit* ambiguity, but also what we have called *implicit* ambiguity–cases in which subjects do not provide evidence of being consciously aware of the ambiguity, but the presence of ambiguity is revealed by the existence of two or more annotators in disagreement. We will address these issues in a future annotation experiment.

4 MEASURING AGREEMENT ON (AMBIGUOUS) ANAPHORIC ANNOTATION

In the discussion above we only gave 'raw' figures of agreement; in this section we move on to the problem of measuring agreement above chance for the annotation of anaphora allowing for explicit ambiguity.

The agreement coefficient which is most widely used in NLP is the one called K by Siegel and Castellan (1988). However, most authors who attempted anaphora annotation pointed out that K is not appropriate for anaphoric annotation. The only sensible choice of ‘label’ in the case of (identity) anaphora are anaphoric chains (Passonneau, 2004); but except when a text is very short, few annotators will catch all mentions of the same discourse entity–most forget to mark a few, which means that agreement as measured with K is always very low. Following Passonneau (2004), we used the coefficient α of Krippendorff (1980) for this purpose, which allows for partial agreement among anaphoric chains. In addition, we developed a new distance metric allowing us to use α to measure agreement when coders are allowed to mark explicit ambiguity.

4.1 Krippendorf's alpha

The α coefficient measures agreement among a set of coders C who assign each of a set of items I to one of a set of distinct and mutually exclusive categories K; for anaphora annotation the coders are the annotators, the items are the markables in the text, and the categories are the emerging anaphoric chains. The coefficient measures the observed disagreement between the coders $\mathrm{D_o}$, and corrects for chance by removing the amount of disagreement expected by chance $\mathrm{D_e}$. The result is subtracted from 1 to yield a final value of agreement.

$$\alpha = 1 - \frac{\mathrm{D_o}}{\mathrm{D_e}}$$

As in the case of K, the higher the value of α, the more agreement there is between the annotators. $\alpha = 1$ means that agreement is complete, and $\alpha = 0$ means that agreement is at chance level.

What makes α particularly appropriate for anaphora annotation is that the categories are not required to be disjoint; instead, they must be ordered according to a DISTANCE METRIC–a function $\mathbf{d}$ from category pairs to real numbers that specifies the amount of dissimilarity between the categories. The distance between a category and itself is always zero, and the less similar two categories are, the larger the distance between them. Table 2 gives the formulas for calculating the observed and expected disagreement for α. The amount of disagreement for each item $i \in I$ is the arithmetic mean of the distances between the pairs of judgments pertaining to it, and the observed disagreement $\mathrm{D_o}$ is the mean of all the item disagreements. The expected disagreement $\mathrm{D_e}$ is the mean of the distances between all the judgment pairs in the data, without regard to items.

$$\begin{aligned}
\mathrm{D_o} &= \frac{1}{\mathbf{ic}(\mathbf{c}-1)} \sum_{i \in I} \sum_{k \in K} \sum_{k' \in K} \mathbf{n}_{ik}\mathbf{n}_{ik'}\mathbf{d}_{kk'} \\
\mathrm{D_e} &= \frac{1}{\mathbf{ic}(\mathbf{ic}-1)} \sum_{k \in K} \sum_{k' \in K} \mathbf{n}_{k}\mathbf{n}_{k'}\mathbf{d}_{kk'}
\end{aligned}$$

$\mathbf{c}$	number of coders
$\mathbf{i}$	number of items
$\mathbf{n}_{ik}$	number of times item i is classified in category k
$\mathbf{n}_{k}$	number of times any item is classified in category k
$\mathbf{d}_{kk'}$	distance between categories k and k'

Table 2: Observed and expected disagreement for α

4.2 Distance measures for anaphora

The distance metric **d** is not part of the general definition of α, because different metrics are appropriate for different types of categories. For anaphora annotation, the most plausible categories are the ANAPHORIC CHAINS: the sets of markables which are mentions of the same discourse entity. Passonneau (2004) proposes a distance metric between anaphoric chains based on the following rationale: two sets are minimally distant when they are identical and maximally distant when they are disjoint; between these extremes, sets that stand in a subset relation are closer (less distant) than ones that merely intersect. This leads to the following distance metric between two sets A and B.

$$\mathbf{d}_{AB}^{\mathit{Passonneau}} = \begin{cases} 0 & \text{if } A = B \\ {}^{1}\!/_{3} & \text{if } A \subset B \text{ or } B \subset A \\ {}^{2}\!/_{3} & \text{if } A \cap B \neq \emptyset \text{, but } A \not\subset B \text{ and } B \not\subset A \\ 1 & \text{if } A \cap B = \emptyset \end{cases}$$

Passonneau's metric is not easy to generalize when ambiguity is allowed. Our generalized measures were based instead on distance metrics commonly used in Information Retrieval that take the size of the anaphoric chain into account, such as Jaccard and Dice (Manning and Schuetze, 1999), the rationale being that the larger the overlap between two anaphoric chains, the better the agreement should be.

$$\begin{aligned}
\mathrm{Jaccard}(A,B) &= \frac{|A \cap B|}{|A \cup B|} \\
\mathrm{Dice}(A,B) &= \frac{2\,|A \cap B|}{|A| + |B|}
\end{aligned}$$

Jaccard and Dice's set comparison metrics were subtracted from 1 in order to get measures of distance that range between zero (minimal distance, identity) and one (maximal distance, disjointness).

$$\begin{aligned}\mathbf{d}_{AB}^{Jaccard} &= 1-\mathrm{Jaccard}(A,B)\\ \mathbf{d}_{AB}^{Dice} &= 1-\mathrm{Dice}(A,B)\end{aligned}$$

The Dice measure always gives a smaller distance than the Jaccard measure, hence Dice always yields a higher agreement coefficient than Jaccard when the other conditions remain constant. The difference between Dice and Jaccard grows with the size of the compared sets.

4.3 Extending α to measure agreement on ambiguity

The distance measures discussed above can be generalized as follows to use α as our measure of agreement in cases in which more than one antecedent has been marked. First of all, we will assume that an ambiguous expression denotes a set of 'normal' interpretations–in the case of anaphora, a set of anaphoric chains. In other words, if *w* is judged as ambiguous, either expressing discourse entity $\{\mathsf{x}_1 \ldots \mathsf{x}_n\}$ or discourse entity $\{\mathsf{y}_1 \ldots \mathsf{y}_n\}$, it will get as a label the set of sets $\{\{\mathsf{x}_1 \ldots \mathsf{x}_n\}, \{\mathsf{y}_1 \ldots \mathsf{y}_n\}\}$. In order to treat all anaphoric expressions uniformly, we use sets of sets to represent the judgments for *all* expressions. Thus, when an anaphoric expression is interpreted as unambiguous, and as a realization of the discourse entity with mentions $\mathsf{x}_1 \ldots \mathsf{x}_n$, it will be assigned as a label the singleton set of sets $\{\{\mathsf{x}_1 \ldots \mathsf{x}_n\}\}$.

Now, intuitions about ambiguity judgments are not always very clear. It probably doesn't make sense to try to arrive at absolute values; but in some cases we at least aim to get reasonable intuitions concerning the relative value of **d** for certain pairs of labels. One case that is clear is that $\mathbf{d}_{AB} = 0$ when both annotators assign the same label to an object, whether that label is unambiguous or ambiguous. (Keep in mind that $\mathbf{d}_{AB}$ measures *disagreement*, not agreement.) It seems equally clear that $\mathbf{d}_{AB} = 1$ when the labels are entirely different – again, whether ambiguous or unambiguous. It also seems clear that just as in the case of unambiguous anaphoric annotation, partial credit should be assigned when there is some overlap between the annotations. One constraints we can impose is that the agreement value for only partially overlapping labels should be less than the value when these labels are identical, yet higher than in the case of completely different labels.

We can define measures of disagreements with the properties above as follows. We begin by introducing generalizations of the Dice and Jaccard measures that work over sets of sets:

$$\begin{aligned}
GJacc(A1,A2) &= \max_m \frac{\sum Jacc(A1_i, m(A1_i))}{|A1| \cup |A2|} \\
GDice(A1,A2) &= \max_m \frac{2 * \sum Dice(A1_i, m(A1_i))}{|A1| + |A2|}
\end{aligned}$$

We can then introduce modified versions of **d** as follows:

$$\begin{aligned}
\mathbf{d}_{AB}^{GeneralizedJaccard} &= 1 - \mathrm{GJacc}(A,B) \\
\mathbf{d}_{AB}^{GeneralizedDice} &= 1 - \mathrm{GDice}(A,B)
\end{aligned}$$

For illustration purposes, values of $\mathbf{d}^{GeneralizedDice}$ for a few examples of coder judgments about the coreference chain to which an anaphoric expression belongs are shown in Table 3. (Remember that with α, we are measuring *dis*agreement, so 0 means perfect agreement.)

	Coder 1	Coder 2	$\mathbf{d}^{GeneralizedDice}$
Identical unambiguous	$\{\{x,y\}\}$	$\{\{x,y\}\}$	0
Identical ambiguous	$\{\{x\},\{y\}\}$	$\{\{x\},\{y\}\}$	0
Overlapping judgments	$\{\{x\},\{y\}\}$	$\{\{x\}\}$	$\frac{1}{3}$

Table 3: Example values of **d** with Generalized Dice

4.4 Agreement on Ambiguous Anaphoric Annotation

The agreement values obtained using α with the generalized distance measures discussed above are shown in Table 4 (first half of the dialogue) and Table 5 (second half). The calculation of α was manipulated under the following three conditions.

Place markables. We calculated the value of α on the entire set of markables (with the exception of three which had data errors), and also on a subset of markables – those that were not place names. Agreement on marking place names was almost perfect: 45 of the 48 place name markables were marked correctly as “place” by all 18 subjects, two were marked correctly by all but one subject, and one was marked correctly by all but two subjects. Place names thus contributed substantially to the agreement among the subjects. Dropping these markables from the analysis resulted in a substantial drop in the value of α across all conditions.

	With place markables		Without place markables	
	Jacc	Dice	Jacc	Dice
No chain	0.64854	0.65558	0.52866	0.53808
Partial	0.63724	0.67044	0.51285	0.55657
Inclusive [−top]	0.61505	0.67920	0.48159	0.56663
Exclusive [−top]	0.58208	0.63524	0.43676	0.50636
Inclusive [+top]	0.56022	0.63195	0.44799	0.53951
Exclusive [+top]	0.51880	0.57826	0.39358	0.46819

Table 4: Agreement with ambiguity: first half of the dialogue

	With place markables		Without place markables	
	Jacc	Dice	Jacc	Dice
No chain	0.67257	0.67633	0.46719	0.47324
Partial	0.65804	0.68181	0.44340	0.48060
Inclusive [−top]	0.65622	0.69167	0.43764	0.49386
Exclusive [−top]	0.63153	0.66141	0.39701	0.44387
Inclusive [+top]	0.59606	0.64296	0.41112	0.47936
Exclusive [+top]	0.55961	0.59989	0.35578	0.41359

Table 5: Agreement with ambiguity: second half of the dialogue

Distance measure. We used the two generalized measures discussed earlier to calculate distance between sets: Jaccard and Dice.[10]

Chain construction. As we report elsewhere, substantial variation in the agreement values can be obtained by making changes to the way anaphoric chains are constructed. We tested the following methods.

NO CHAIN: only the immediate antecedents of an anaphoric expression were considered, instead of building an anaphoric chain.

PARTIAL CHAIN: a markable's chain included only phrase markables which occurred in the dialogue before the markable in question (as well as all discourse markables).

FULL CHAIN: chains were constructed by looking upward and then back down, including all phrase markables which occurred in the dialogue either before or after the markable in question (as well as the markable itself, and all discourse markables).

We used two separate versions of the full chain condition: in the [+top] version we associate the top of a chain with the chain itself, whereas in the [−top] version we associate the top of a chain with its original category label, "place" or "none".

Passonneau (2004) observed that in the calculation of observed agreement, two full chains always intersect because they include the current item. Passonneau suggests to prevent this by excluding the current item from the chain for the purpose of calculating the observed agreement. We performed the calculation both ways – the inclusive condition includes the current item, while the exclusive condition excludes it.

The four ways of calculating α for full chains, plus the no chain and partial chain condition, yield the six chain conditions in Tables 4 and 5. Other things being equal, Dice yields a higher agreement than Jaccard.

The exclusive chain conditions always give lower agreement values than the corresponding inclusive chain conditions, because excluding the current item reduces observed agreement without affecting expected agreement (there is no "current item" in the calculation of expected agreement).

[10] Passonneau's measure cannot easily be generalized to multiple sets. For the nominal categories "place" and "none" we assign a distance of zero between the category and itself, and of one between a nominal category and any other category.

	With place markables	Without place markables
No chain	0.62773	0.50066
Partial	0.56175	0.41255
Full [−top]	0.47937	0.30260
Full [+top]	0.39328	0.23678

Table 6: Experiment 1a Kappa (π) values with ambiguity

	With place markables	Without place markables
No chain	0.66201	0.44997
Partial	0.59403	0.34286
Full [−top]	0.55201	0.27330
Full [+top]	0.45441	0.20687

Table 7: Experiment 1b Kappa (π) values with ambiguity

The [−top] conditions tended to result in a higher agreement value than the corresponding [+top] conditions because the tops of the chains retained their “place” and “none” labels; not surprisingly, the effect was less pronounced when place markables were excluded from the analysis. Inclusive [−top] was the only full chain condition which gave α values comparable to the partial chain and no chain conditions. For each of the four selections of markables, the highest α value was given by the Inclusive [−top] chain with Dice measure.

For comparison purposes, we also report in Tables 6 and 7 the values obtained with K (as defined by Siegel and Castellan (1988)) instead of α–i.e., by not giving partial ’credit’ to cases of partial overlap between partial chains. No difference is found for the no-chain condition, as expected, but for all other conditions the values of agreement are systematically lower than those obtained with α.

5 DISCUSSION

In summary, the main contributions of this work so far have been to further develop the methodology for annotating anaphoric relations by (i) testing methods for annotating some types of anaphoric ambiguity, and (ii) developing techniques for measuring agreement on this type of annotation. Our preliminary analysis revealed the need in future experiments to introduce methods to mark the discourse-new / discourse-old ambiguity, and to clarify the difference between ambiguity and reference to multiple objects. More seriously, our studies found that with our current instructions, in most cases annotators are not aware of the ambiguity, so that ambiguity is only revealed when comparing the annotations, rather than being explicitly marked. While this is not a problem when the goal is simply that of identifying problematic cases of anaphoric reference, the implications of this finding from the point of view of developing a reliable scheme for anaphoric annotation still need to be considered.

Our future work will include further developments of the annotation methodology, including also more advanced methods for marking discourse deixis, and of the methodology for measuring agreement with ambiguous annotations.

ACKNOWLEDGMENTS

This work was in part supported by EPSRC project GR/S76434/01, ARRAU. We wish to thank Tony Sanford, Patrick Sturt, Ruth Filik, Harald Clahsen, Sonja Eisenbeiss, and Claudia Felser.

References

Buitelaar, P. (1998). *CoreLex : Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University.

Byron, D. (2003). Annotation of pronouns and their antecedents: A comparison of two domains. Technical Report 703, University of Rochester, Computer Science Department.

Dale, R. (1992). *Generating Referring Expressions*. The MIT Press, Cambridge, MA.

Eckert, M. and Strube, M. (2001). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*.

Gross, D., Allen, J., and Traum, D. (1993). The TRAINS 91 dialogues. TRAINS Technical Note 92-1, Computer Science Dept. University of Rochester.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, **69**(2), 274–307.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst.

Hirschman, L. (1998). MUC-7 coreference task definition, version 3.0. In N. Chinchor, editor, *In Proc. of the 7th Message Understanding Conference*. Available at `http://www.muc.saic.com/proceedings/muc_7_toc.html`.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. D. Reidel, Dordrecht.

Krippendorff, K. (1980). *Content Analysis: An introduction to its Methodology*. Sage Publications.

Kurtzman, H. S. and MacDonald, M. C. (1993). Resolution of quantifier scope ambiguities. *Cognition*, **48**, 243–279.

Manning, C. D. and Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Müller, C. and Strube, M. (2003). Multi-level annotation in MMAX. In *Proc. of the 4th SIGDIAL*, pages 198–207.

Palmer, M., Dang, H., and Fellbaum, C. (2005). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

Passonneau, R. J. (1997). Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript.

Passonneau, R. J. (2004). Computing reliability for coreference annotation. In *Proc. of LREC*, Lisbon.

Pinkal, M. (1995). *Logic and Lexicon*. D. Reidel, Dordrecht.

Poesio, M. (1996). Semantic ambiguity and perceived ambiguity. In K. van Deemter and S. Peters, editors, *Semantic Ambiguity and Underspecification*, chapter 8, pages 159–201. CSLI, Stanford, CA.

Poesio, M. (2004). The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL*, Boston.

Poesio, M. (To appear). *Incrementality and Underspecification in Semantic Interpretation*. Lecture Notes. CSLI, Stanford, CA.

Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, **24**(2), 183–216.

Poesio, M., Bruneseaux, F., and Romary, L. (1999). The MATE meta-scheme for coreference in dialogues in multiple languages. In M. Walker, editor, *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 65–74.

Poesio, M., Reyle, U., and Stevenson, R. (To appear). Justified sloppiness in anaphoric reference. In H. Bunt and R. Muskens, editors, *Computing Meaning 3*. Kluwer.

Prince, E. F. (1992). The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund-raising text*, pages 295–325. John Benjamins.

Raskin, V. (1985). *Semantic Mechanisms of Humor*. D. Reidel, Dordrecht and Boston.

Rosenberg, A. and Binkowski, E. (2004). Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proc. of NAACL*, volume Short papers.

Sidner, C. L. (1979). *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition.

Su, S. P. (1994). *Lexical Ambiguity in Poetry*. Longman, London.

van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, **26**(4), 629–637. Squib.

Webber, B. L. (1979). *A Formal Approach to Discourse Anaphora*. Garland, New York.