The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account

Massimo Poesio and Ron Artstein
University of Essex,
Language and Computation Group / Department of Computer Science
United Kingdom

Abstract

We report the results of a study of the reliability of anaphoric annotation which (i) involved a substantial number of naive subjects, (ii) used Krippendorff's α instead of K to measure agreement, as recently proposed by Passonneau, and (iii) allowed annotators to mark anaphoric expressions as ambiguous.

1 INTRODUCTION

We tackle three limitations with the current state of the art in the annotation of anaphoric relations. The first problem is the lack of a truly systematic study of agreement on anaphoric annotation in the literature: none of the studies we are aware of (Hirschman, 1998; Poesio and Vieira, 1998; Byron, 2003; Poesio, 2004) is completely satisfactory, either because only a small number of coders was involved, or because agreement beyond chance couldn't be assessed for lack of an appropriate statistic, a situation recently corrected by Passonneau (2004). The second limitation, which is particularly serious when working on dialogue, is our still limited understanding of the degree of agreement on references to abstract objects, as in discourse deixis (Webber, 1991; Eckert and Strube, 2001).

The third shortcoming is a problem that affects all types of semantic annotation. In all annotation studies we are aware of,[1] 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. Consider the dialogue excerpt in (1):[2] it's not clear to us (nor was to our annotators, as we'll see below) whether the demonstrative *that* in utterance unit 18.1 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)
```
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
```

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

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

[2] 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`.

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 76–83,
Ann Arbor, June 2005.

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.

The paper is organized as follows. We first briefly review previous work on anaphoric annotation and on reliability indices. We then discuss our experiment with anaphoric annotation, and its results. Finally, we discuss the implications of this work.

2 ANNOTATING ANAPHORA

It is not our goal at this stage to propose a new scheme for annotating anaphora. For this study we simply developed a coding manual for the purposes of our experiment, broadly based on the approach adopted in MATE (Poesio et al., 1999) and GNOME (Poesio, 2004), but introducing new types of annotation (ambiguous anaphora, and a simple form of discourse deixis) while simplifying other aspects (e.g., by not annotating bridging references).

The task of 'anaphoric annotation' discussed here is related, although different from, the task of annotating 'coreference' in the sense of the so-called MUCSS scheme for the MUC-7 initiative (Hirschman, 1998). This scheme, while often criticized, is still 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 expressed by the scheme, ident, conflates a number of relations that semanticists view as distinct: besides COREFERENCE proper, there are IDENTITY ANAPHORA, 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 / psychological perspective, and, from a computational perspective, for the evaluation of anaphora resolution and referring expressions generation. The goal is to annotate the *discourse model* resulting from the interpretation of a text, in the sense both of (Webber, 1979) and of dynamic theories of anaphora (Kamp and Reyle, 1993). In order to do this, annotators must first of all identify the noun phrases which either introduce new discourse entities (discourse-new (Prince, 1992)) or are mentions of previously introduced ones (discourse-old), ignoring those that are used predicatively. Secondly, annotators have to specify which discourse entities have the same interpretation. Given that the characterization of such discourse models is usually considered part of the area of the semantics of anaphora, and that the relations to be annotated include relations other than Sidner's (1979) COSPECIFICATION, we will use the term ANNOTATION OF ANAPHORA for this task (Poesio, 2004), but the reader should keep in mind that we are not concerned only with nominal expressions which are lexically anaphoric.

3 MEASURING AGREEMENT ON ANAPHORIC ANNOTATION

The agreement coefficient which is most widely used in NLP is the one called K by Siegel and Castellan (1988). Howewer, 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.[3]

3.1 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 D_o, and corrects for

[3] We also tried a few variants of α, but these differed from α only in the third to fifth significant digit, well below any of the other variables that affected agreement. In the interest of space we only report here the results obtained with α.

chance by removing the amount of disagreement expected by chance D_e. The result is subtracted from 1 to yield a final value of agreement.

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

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 **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 1 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 agreement is the mean of all the item disagreements. The expected disagreement is the mean of the distances between all the judgment pairs in the data, without regard to items.

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

$$D_e = \frac{1}{\mathbf{ic}(\mathbf{ic}-1)} \sum_{k \in K} \sum_{k' \in K} \mathbf{n}_k \mathbf{n}_{k'} \mathbf{d}_{kk'}$$

c	number of coders
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 1: Observed and expected disagreement for α

3.2 Distance measures

The distance metric is not part of the general definition of α, because different metrics are appropriate for different types of categories. For anaphora annotation, the 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.

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

We also tested 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. 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).

$$\mathbf{d}_{AB} = 1 - \frac{|A \cap B|}{|A \cup B|} \qquad \text{(Jaccard)}$$

$$\mathbf{d}_{AB} = 1 - \frac{2|A \cap B|}{|A| + |B|} \qquad \text{(Dice)}$$

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. Obviously, the Passonneau measure is not sensitive to the size of these sets.

3.3 Computing the anaphoric chains

Another factor that affects the value of the agreement coefficient–in fact, arguably the most important factor–is the method used for constructing from the raw annotation data the 'labels' used for agreement computation, i.e., the anaphoric chains. We experimented with a number of methods. However, since the raw data are highly dependent on the annotation scheme, we will postpone discussing our chain construction methods until after we have described our experimental setup and annotation scheme. We will also discuss there how comparisons are made when an ambiguity is marked.

4 THE ANNOTATION STUDY

4.1 The Experimental Setup

Materials. 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).[4]

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.

Procedure. The subjects performed the experiment together in one lab, each working on a separate computer. 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 the dialogue (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.

4.2 The Annotation Scheme

MMAX 2 allows for multiple types of markables; markables at the phrase, utterance, and turn levels were defined before the experiment. 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 mark each of them with one of four attributes: "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 Avon, Bath, Corning, Dansville, and Elmira, explicitly mentioned by name; or "none" if it did not fit any of the above criteria, for instance if it referred to a novel object or was not a referential noun phrase. (We included the attribute "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.) For markables designated as "phrase" or "segment" subjects were instructed to set a pointer to the antecedent, a markable at the phrase or turn level. Subjects were instructed to set more than one pointer in case of ambiguous reference. Markables which were not given an attribute or which were marked as "phrase" or "segment" but did not have an antecedent specified were considered to be 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 creating an association between them. 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. In a separate experiment we tested a more complicated annotation scheme which allowed a more fine-grained marking of reference to discourse segments.

4.3 Computing anaphoric chains

The raw annotation data were processed using custom-written Perl scripts to generate coreference chains and calculate reliability statistics.

The core of Passonneau's proposal (Passonneau, 2004) is her method for generating the set of dis-

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

tinct and mutually exclusive categories required by α out of the raw data of anaphoric annotation. Considering as categories the immediate antecedents would mean a disagreement every time two annotators mark different members of an anaphoric chain as antecedents, while agreeing that these different antecedents are part of the same chain. Passonneau proposes the better solution to view the emerging anaphoric chains themselves as the categories. And in a scheme where anaphoric reference is unambiguous, these chains are equivalence classes of markables. But we have a problem: since our annotation scheme allows for multiple pointers, these chains take on various shapes and forms.

Our solution is to associate each markable *m* with the set of markables obtained by following the chain of pointers from *m*, and then following the pointers backwards from the resulting set. The rationale for this method is as follows. Two pointers *to* a single markable never signify ambiguity: if *B* points to *A* and *C* points to *A* then *B* and *C* are cospecificational; we thus have to follow the links up and then back down. However, two pointers *from* a single markable may signify ambiguity, so we should not follow an up-link from a markable that we arrived at via a down-link. The net result is that an unambiguous markable is associated with the set of all markables that are cospecificational with it on one of their readings; an ambiguous markable is associated with the set of all markables that are cospecificational with at least one of its readings. (See figure 1.)

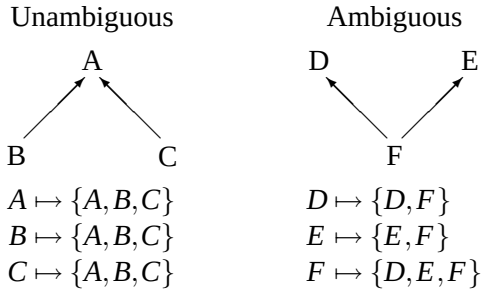

Figure 1: Anaphoric chains

This method of chain construction also allows to resolve apparent discrepancies between reference to phrase-level and turn-level markables. Take for example the snippet below: many annotators marked a pointer from the demonstrative *that* in utterance unit 4.2 to turn 3; as for *that* in utterance unit 4.3, some marked a pointer to the previous *that*, while others marked a pointer directly to turn 3.

(2)
```
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
```

In this case, not only do the annotators mark different direct antecedents for the second *that*; they even use different attributes–"phrase" when pointing to a phrase antecedent and "segment" when pointing to a turn. Our method of chain construction associates both of these markings with the same set of three markables – the two *that* phrases and turn 3 – capturing the fact that the two markings are in agreement.[5]

4.4 Taking ambiguity into account

The cleanest way to deal with ambiguity would be to consider each item for which more than one antecedent is marked as denoting a set of interpretations, i.e., a set of anaphoric chains (Poesio, 1996), and to develop methods for comparing such sets of sets of markables. However, while our instructions to the annotators were to use multiple pointers for ambiguity, they only followed these instructions for phrase references; when indicating the referents of discourse deixis, they often used multiple pointers to indicate that more than one turn had contributed to the development of a plan. So, for this experiment, we simply used as the interpretation of markables marked as ambiguous the union of the constituent interpretations. E.g., a markable E marked as pointing both to antecedent A, belonging to anaphoric chain {A,B}, and to antecedent C, belonging to anaphoric chain {C,D}, would be treated by our scripts as being interpreted as referring to anaphoric chain {A,B,C,D}.

5 RESULTS

5.1 Agreement on category labels

The following table reports for each of the four categories the number of cases (in the first half) in which

[5]It would be preferable, of course, to get the annotators to mark such configurations in a uniform way; this however would require much more extensive training of the subjects, as well as support which is currently unavailable from the annotation tool for tracking chains of pointers.

a good number (18, 17, 16) annotators agreed on a particular label–phrase, segment, place, or none–or no annotators assigned a particular label to a markable. (The figures for the second half are similar.)

Number of judgments	18	17	16	0
phrase	10	3	1	30
segment			1	52
place	16	1	1	54
none	10	5	1	29

Table 2: Cases of good agreement on categories

In other words, in 49 cases out of 72 at least 16 annotators agreed on a label.

5.2 Explicitly annotated ambiguity, and its impact on agreement

Next, we attempted to get an idea of the amount of *explicit* ambiguity–i.e., the cases in which coders marked multiple antecedents–and the impact on reliability resulting by allowing them to do this. In the first half, 15 markables out of 72 (20.8%) were marked as explicitly ambiguous by at least one annotator, for a total of 55 explicit ambiguity markings (45 phrase references, 10 segment references); in the second, 8/76, 10.5% (21 judgments of ambiguity in total). The impact of these cases on agreement can be estimated by comparing the values of K and α on the antecedents only, before the construction of cospecification chains. Recall that the difference between the coefficients is that K does not allow for partial disagreement while α gives it some credit. Thus if one subject marks markable *A* as antecedent of an expression, while a second subject marks markables *A* and *B*, K will register a disagreement while α will register partial agreement. Table 3 compares the values of K and α, computed separately for each half of the dialogue, first with all the markables, then by excluding "place" markables (agreement on marking place names was almost perfect, contributing substantially to overall agreement). The value of α is somewhat higher than that of K, across all conditions.

5.3 Agreement on anaphora

Finally, we come to the agreement values obtained by using α to compare anaphoric chains computed as discussed above. Table 4 gives the value of α for the first half (the figures for the second half are similar). The calculation of α was manipulated under the following three conditions.

		With place	Without place
First Half	K	0.62773	0.50066
	α	0.65615	0.53875
Second Half	K	0.66201	0.44997
	α	0.67736	0.47490

The coefficient reported here as K is the one called K by Siegel and Castellan (1988).
The value of α is calculated using Passonneau's distance metric; for other distance metrics, see table 4.

Table 3: Comparing K and α

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.

Distance measure. We used the three measures discussed earlier to calculate distance between sets: Passonneau, Jaccard, and Dice.[6]

Chain construction. Substantial variation in the agreement values can be obtained by making changes to the way we construct anaphoric chains. 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 dia-

[6]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		
	Pass	Jacc	Dice	Pass	Jacc	Dice
No chain	0.65615	0.64854	0.65558	0.53875	0.52866	0.53808
Partial	0.67164	0.65052	0.67667	0.55747	0.53017	0.56477
Inclusive [−top]	0.65380	0.64194	0.69115	0.53134	0.51693	0.58237
Exclusive [−top]	0.62987	0.60374	0.64450	0.49839	0.46479	0.51830
Inclusive [+top]	0.60193	0.58483	0.64294	0.49907	0.47894	0.55336
Exclusive [+top]	0.57440	0.53838	0.58662	0.46225	0.41766	0.47839

Table 4: Values of α for the first half of dialogue 3.2

logue 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 Table 4. Other things being equal, Dice yields a higher agreement than Jaccard; considering both halves of the dialogue, the Passonneau measure always yielded a higher agreement that Jaccard, while being higher than Dice in 10 of the 24 conditions, and lower in the remaining 14 conditions.

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).

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.

5.4 Qualitative Analysis

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*,[7] all other cases of disagreement reflect a genuine ambiguity, as opposed to differences in the application of subjective categories.[8]

Lack of space prevents a full discussion of the data, but some of the main points can already be made with reference to the part of the dialogue in (2), repeated with additional context in (3).

[7] 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%.

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

(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' reference. 9 coders marked the pronoun as 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. This example shows that when trying to develop methods to identify ambiguous cases it is important to consider not only the cases of *explicit* ambiguity, but also so-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 (Poesio, 1996).

6 DISCUSSION

In summary, the main contributions of this work so far has been (i) to further develop the methodology for annotating anaphoric relations and measuring the reliability of this type of annotation, adopting ideas from Passonneau and taking ambiguity into account; and (ii) to run the most extensive study of reliability on anaphoric annotation todate, showing the impact of such choices. Our future work includes further developments of the methodology for measuring agreement with ambiguous annotations and for annotating discourse deictic references.

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

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

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

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

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

L. Hirschman. 1998. MUC-7 coreference task definition, version 3.0. In N. Chinchor, editor, *In Proc. of the 7th Message Understanding Conference*.

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

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

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

C. Müller and M. Strube. 2003. Multi-level annotation in MMAX. In *Proc. of the 4th SIGDIAL*.

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

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

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

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

M. Poesio, F. Bruneseaux, and L. Romary. 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.

M. Poesio. 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.

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

E. F. Prince. 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.

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

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

S. Siegel and N. J. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill.

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

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

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