Identifying reference to abstract objects in dialogue

Ron Artstein and Massimo Poesio
Department of Computer Science
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom
artstein|poesio [at] essex.ac.uk

Abstract

In two experiments, many annotators marked antecedents for discourse deixis as unconstrained regions of text. The experiments show that annotators do converge on the identity of these text regions, though much of what they do can be captured by a simple model. Demonstrative pronouns are more likely than definite descriptions to be marked with discourse antecedents. We suggest that our methodology is suitable for the systematic study of discourse deixis.

1 Introduction

This paper describes two experiments that used corpus annotation to characterize discourse deixis (Webber, 1991)—an anaphoric relation in dialogue, where the reference of an anaphoric expression is present in the preceding text but not in the form of an explicit antecedent. An example of such a relation can be seen in the interpretation of the demonstrative pronoun *that* in the following snippet, taken from dialogue 2.2 of the TRAINS-91 corpus (Gross et al., 1993).[1]

(1) 7.3 : so we ship one
7.4 : boxcar
7.5 : of oranges to Elmira
7.6 : and that takes another 2 hours

The reference of *that* clearly depends on the preceding text, and in this sense the pronoun is an anaphor. The meaning of *that* in this context can perhaps be expressed with a nominalization such as *the shipping of one boxcar of oranges to Elmira*. Such a nominalization is not present in the text—but something very close to it is. This paper addresses the problem of how the appropriate antecedent can be identified through corpus annotation.

Previous work on annotating discourse-deictic relations has achieved reliability at the cost of severe restrictions on the annotation (Byron, 2002; Eckert and Strube, 2000; Navarretta, 2000). However, there is a need for empirical work to determine the degree of objectivity concerning the identification of specific references to abstract objects, even if only to conclude that such references are interpreted so subjectively that it wouldn't make sense for a system to resolve them. The experiments reported here were designed to assess the feasibility of identifying such anaphoric relations using a fairly unconstrained annotation format and a large number of linguistically naive annotators. We exchanged the highly knowledgeable opinions (and prejudices) of experts with the collective wisdom of many speakers, looking for interesting patterns that would emerge.

The references of the anaphors in question are often abstract, and do not necessarily correspond to any particular phrase or clause in the text. It is often possible to characterize an abstract referent with a textual description, as we did for the reference of the anaphor in example (1); however, we have no systematic way to compare characterizations by different annotators. In the absence of an explicit representation of all the potential referents, we chose to have our annotators point out the required antecedents by marking unconstrained regions in the text of the dialogue; this allowed comparing the annotations while retaining a high degree of precision.

2 The TRAINS dialogues

The dialogues annotated in the experiments come from the first edition of the TRAINS corpus collected at the University of Rochester in 1991 (Gross et al., 1993). This corpus consists of tran-

[1] The TRAINS-91 dialogue transcripts are available at ftp://ftp.cs.rochester.edu/pub/papers/ai/92.tn1.trains_91_dialogues.txt

brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, pages 56–63.

scripts of dialogues between two humans. One of the humans plays the 'manager' of a railway company, who needs to develop a plan to deliver specific goods at particular stations by a given deadline. The other participant in the dialogue plays a 'system', whose role is to provide the manager with required information such as journey times and equipment availability. The corpus consists of sixteen dialogues performed by eight different 'managers'—each manager has a short dialogue with a simple problem to become familiarized with the task, and a longer dialogue with a more complicated problem to solve. The 'system' in all sixteen dialogues is played by the same person.

The dialogues thus have a quite limited domain. The participants refer often to objects in the 'TRAINS world' such as engines, cars, stations, and commodities; they talk about routes, distances and times, and about different possibilities for moving the objects around. They formulate plans and identify conflicts between them. Because the goals of the dialogues are constrained, the range of abstract objects that are discussed in the dialogues is also quite limited. This is an advantage for the present study because it makes the (unconstrained) responses of the annotators fairly tractable and interpretable.

3 Annotation

The coding manual used in the experiments was based on the approach developed in the projects MATE (Poesio et al., 1999) and GNOME (Poesio, 2004). The task and instructions were simplified by eliminating the annotation of bridging references; on the other hand, we added instructions for marking multiple antecedents for ambiguous anaphoric expressions, and for marking text regions to represent abstract antecedents.

The dialogue transcripts were annotated on a computer, using the MMAX 2 annotation tool (Müller and Strube, 2003).[2] This tool uses an XML format which allows the definition of multiple levels of *markables* on top of a base text, for example phrase markables and utterance markables. We used the tool's project wizard to create the experimental texts from the plain-text transcripts and to automatically create utterance-level markables; we then manually defined the phrase-level markables, which included all the noun phrases in the text (except temporal ones). The phrase-level and utterance-level markables were the same for all the experiment participants, except for very few cases where in the course of annotation a participant inadvertently deleted or redefined a markable (this was due to a limitation of the tool, which does not afford the possibility of fixing the identity of markables while marking their attributes; the tool does make sure, however, that participants cannot alter the base text).

The participants entered their annotations using the graphical interface of MMAX 2. Their task was to determine, for all the predefined phrase-level markables in the text, whether they were anaphoric, and to identify antecedents for the anaphoric ones; antecedents were marked by creating *pointers* from an anaphoric markable to another markable representing the antecedent. If the antecedent was mentioned previously by an expression which was a phrase-level markable, a pointer was set from the anaphoric markable to the antecedent markable. If the antecedent was *not* mentioned previously by a phrase-level markable, then a text region was marked as the antecedent. The marking of text regions was done somewhat differently in the two experiments. In experiment 1, participants defined markables on a separate level, the *segment level* (hence the term "segment antecedent"); they were thus able to mark arbitrary regions of text to represent abstract antecedents (even discontinuous regions). This allowed the annotators to make very fine-grained distinctions. For example, a reasonable interpretation of the following part of dialogue 2.2 gives slightly different referents to the pronouns *that* in utterances 3.6 and 3.7: the pronoun *that* in 3.6 refers to getting the boxcar and engine to Corning, while the pronoun *that* in 3.7 refers to getting the boxcar and engine to Corning *from Elmira*.

(2) 3.1 M: so
3.2 : essentially we have to
3.3 : ... again get the boxcar
3.4 : and engine
3.5 : to Corning
3.6 : so the fastest way to do that is from Elmira
3.7 : so we'll do that

Indeed, one of our annotators captured this distinction by pointing the first pronoun to the text *get the boxcar and engine to Corning* while pointing the second pronoun to the text *get the boxcar and*

[2] http://mmax.eml-research.de/

engine to Corning from Elmira (note that the latter is a discontinuous portion of text, which also does not correspond to any syntactic constituent). However, most of the experiment participants did not make such fine-grained distinctions, and the need to define segment markables caused difficulties for some of the participants in the interaction with the software. Therefore in experiment 2 we chose a simpler design in which participants did not define new markables, but rather marked segment antecedents with multiple pointers to individual utterances: segment antecedents were not collections of words, but collections of utterances. This coarser marking of segment antecedents simplified the annotation procedure considerably.

The annotated dialogues (in XML format) were processed with custom-built perl scripts to extract the references to text regions and present them in a form suitable for analysis. Part of this processing involved propagation of these references down the coreference chains. This was needed because sometimes the same abstract object is referred to more than once in the dialogue. For example, in the following snippet from dialogue 2.2, the pronoun *that* in utterance 30.1 may refer to the same plan as the pronoun *it* in utterance 29.2.

(3) 29.1 M: mkay
29.2 : and how long would it take
30.1 S: that would take
30.2 : um
30.3 : ... six hours from .. Elmira

The annotation instructions specified that if the two markables (*it* and *that*) refer to the same object, then the first markable (*it*) should be marked as the antecedent of the second markable (*that*) regardless of whether the referent is concrete or abstract. For the purpose of this study we are interested in identifying all the references to the kind of objects represented by segment antecedents, and therefore for the purpose of analysis we propagated references to segment antecedents down the chains.

4 Experiment 1

This experiment tested the feasibility of marking text regions to represent abstract antecedents, using a large number of naive annotators; it was based on an earlier pilot which showed that inexperienced participants can be trained quickly to master enough of the MMAX 2 software to allow for reasonable annotation performance.

4.1 Experimental setup

Materials Dialogue 2.2 from the TRAINS-91 corpus; dialogue 2.1 was used for training.

Participants Twenty paid undergraduates, native speakers of English, without any previous training in corpus annotation (except one who had previously participated in a similar experiment; subsequent clustering to identify outliers failed to distinguish this participant from the others).

Procedure The participants 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: participants were given the annotation manual and a map of the 'TRAINS world' and taught how to use the software, and then annotated the training text together. After the break, the participants started annotating the experimental dialogue. The second session took place five days later, and each participant continued from the point they had stopped on the previous day. Nineteen of the twenty participants completed the annotation, and continued on to annotate a newswire text as part of a separate experiment.

4.2 Results

Of the 181 phrase markables in the dialogue, 35 were annotated with a segment antecedent by three or more annotators. We chose to ignore the annotations on all markables which were given a segment antecedent by just one or two annotators, as it appears that with 20 annotators in total, such rare annotations are most likely to be errors: of the 26 markables which were identified by only one annotator, all but one appear to be in error, and of the 12 markables identified by just two annotators, at least 6 appear to be in error. The large number of singular annotations is partly due to antecedent propagation: for example, one participant linked the ten occurrences of *orange juice* and *the orange juice* in an anaphoric chain, and marked the top of the chain with a segment antecedent (annotator error); because of antecedent propagation, all ten markables appear to have a segment antecedent—but only by one annotator. A total of eight annotators contributed such singular annotations; thus, the errors do not appear to come from particular annotators who misunderstood the dialogue or in-

structions, but rather look like arbitrary mistakes.

Agreement By and large, annotators seemed to agree with one another on the identity of the segment antecedents they had marked. It is not clear what is the best way to measure the amount of such agreement. One simple measure is to check what percentage of annotators formed the most common choice for each markable. As an example we can look at the following bit of dialogue.

(4) 3.6 : so the fastest way to do that is from Elmira
3.7 : so we'll do that
⋮
7.3 : so we ship one
7.4 : boxcar
7.5 : of oranges to Elmira
7.6 : and that takes another 2 hours

Ten annotators marked segment antecedents for the pronoun *that* in utterance 7.6, and their chosen antecedents are shown in the following table.

Antecedent	*N*
(3.6) the … that (3.7)	1
(7.3) so … Elmira (7.5)	3
(7.3) we … Elmira (7.5)	2
(7.3) ship … Elmira (7.5)	3
(7.3) one … Elmira (7.5)	1

The most commonly chosen word for the beginning of the antecedent was either *so* or *ship*, each chosen by 3 annotators (30%); the most common choice for the end of the antecedent was *Elmira*, agreed upon by 9 annotators (90%). Averaging these percentages over the 16 most readily identifiable anaphors (those given segment antecedents by 8–12 annotators), we found that 42% of the time coders agreed with the most popular choice for the beginning of an antecedent, and 64% of the time they agreed with the most popular choice for the end. While simplistic, this measure seems appropriate for showing that agreement was higher on where the segments ended than on where they began.

One problem with the above measure is that it fails to take into account the fact that the words *so, we, ship,* and *one* in utterance 7.3 are very close, and that the antecedents that begin with these words overlap to a substantial extent. An anonymous reviewer suggested using measures from topic segmentation such as P_k (Beeferman et al., 1999) and WindowDiff (Pevzner and Hearst, 2002); however, it is not clear to us how to adapt these measures to multiple coders, and to a situation where only small segments are selected, rather than a segmentation of the whole text. Another possibility is to use Krippendorff's α (Krippendorff, 1980; Krippendorff, 2004), a chance-corrected coefficient that allows various distance metrics between the coded categories. Alpha measures the *observed distance* D_o, which is the mean distance between all pairs of judgments that pertain to the same markable, and the *expected distance* D_e, which is the mean distance between all pairs of judgments without regard to markables; alpha is then defined as a coefficient which ranges from -1 to 1, with 1 signifying perfect agreement ($D_o = 0$), and 0 signifying chance agreement ($D_o = D_e$).

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

Previous work has used α to calculate agreement on anaphoric chains, treating each anaphoric chain as a set of markables and using measures of set differences as distances between the chains (Passonneau, 2004; Poesio and Artstein, 2005a; Poesio and Artstein, 2005b). A similar approach treats segment antecedents as sets of words; we calculated alpha values for the 16 most readily identifiable anaphors using three distance metrics – Jaccard, Dice, and Passonneau.

	Jaccard	Dice	Passonneau
D_o	0.53	0.43	0.43
D_e	0.95	0.94	0.94
α	0.45	0.55	0.55

These measures show a fair amount of overlap between the chosen segment antecedents, though not close to perfect. It is interesting to note that the expected distance D_e is close to maximal (unity): the reason for this is that there is little overlap between the segment antecedents of different anaphors—we do not find many instances of multiple references to the same abstract object (represented by a text region). Therefore α pretty much reflects the observed agreement ($1 - D_o$), as there is little overlap expected by chance.

Treating antecedents as sets of words does not allow us to see easily where the differences between the annotators lie. We can treat beginnings

and endings of words separately by using the interval version of Krippendorff's α, using individual word indices as a linear scale. For a particular markable, the observed distance is the sum of the squares of the distances between all the pairs of words chosen as antecedent beginnings or ends (this is equivalent to twice the variance about the mean, σ^2); the overall observed distance D_o is the sum of observed distances for all markables. Calculated this way for the 16 most readily identifiable anaphors, the observed distance of the beginnings of antecedents is about 2.5 times the observed distance of the ends of antecedents, confirming our previous observation that agreement on antecedent beginnings is lower than on antecedent ends. The expected distance D_e is the sum of the squares of the distances between all the pairs of words chosen as antecedents for any markable. This gives α values of 0.998 for the beginnings of antecedents and 0.999 for the ends of antecedents, which looks like very high agreement. The reason for this high value of α is an extremely high expected distance D_e, caused by the fact that the segment antecedents are spread over the entire dialogue (1421 words), whereas the segment antecedents of each particular markable tend to be in the same vicinity. The high value of α tells us that annotators are performing much better than choosing antecedent starting and ending points at random from all over the dialogue; this is to be expected, given that segment antecedents tend to be close to the anaphors (Passonneau, 1993).

Since we know that segment antecedents tend to be close to the anaphors, we can try an alternative model for chance agreement: assume that antecedents are always marked a fixed distance from their anaphors. This would associate each antecedent beginning or end with its distance from the beginning of the anaphor. The observed distance D_o remains as before, since for each anaphor all the antecedent beginnings and ends are changed by a constant. The expected distance D_e, however, is lowered considerably, since we have factored out the spreading of anaphors over the dialogue. Calculated this way, we get an α of 0.17 for the beginnings of antecedents and 0.12 for the ends of antecedents. This is extremely low: the annotators performed only 10–20% better than picking random points in relation to the anaphor! This low number is partly because interval α, like any measure of variance, takes squares of distances and is thus very sensitive to outliers. The 16 most readily identifiable anaphors comprise 155 individual annotations. In one of these annotations, the beginning and end of the antecedent lie more than 3 standard deviations away from the mean for the anaphor's antecedents; removing this single outlier brings α up to 0.21 for segment beginnings and 0.25 for segment ends. Removing six more data points where either the beginning or end of the antecedent lie 2.5–3 standard deviations away from the mean brings α up to 0.25 and 0.40, and removing an additional nine data points which lie 2–2.5 standard deviations away from the mean brings α up to 0.35 and 0.65. This shows that the extremely low value of α is the result of a small number of outliers, although even with those outliers removed agreement is far from perfect: a very primitive model of just picking an antecedent which is a fixed distance from the referring anaphor (with some random variation) accounts for much of what the annotators are doing. This could be either because the annotators or the annotation procedure are not very good, or because such a primitive model is fairly good at capturing segment antecedents.

The difference between α values for segment beginnings and ends appears to rise as we remove outliers. However, this is probably not meaningful, since this difference varies greatly depending on the cutoff point for outliers and on the minimum number of annotations a markable needs to receive in order to be considered in the comparison (we did not perform significance tests; see Krippendorff (2004) for the difficulties in calculating confidence intervals for α). The failure to show a difference in chance-corrected agreement for segment beginnings and endings means that the primitive model of a fixed distance from the anaphor is about equally good at describing the beginnings of segment antecedents and their ends; the higher agreement on segment endings is the result of lower variance around the fixed distance.

Demonstratives The annotations revealed an overwhelming preference to assign segment antecedents to demonstratives. With the exception of one instance of *that*, all the demonstrative pronouns were identified as referring to segment antecedents by at least three annotators, among them 20 instances of *that*, 4 instances of *this* and 2 instances of *those*. In contrast, only 2 of the 28 instances of the pronoun *it* were marked with a segment antecedent by three or more annotators.

These two *it* pronouns were marked by just four annotators each, and the segment interpretation of these pronouns is clearly not the only possible one. The first is the pronoun *it* in utterance 13.3: a few annotators marked it as referring to a text region containing utterance 13.1, presumably intending the action of moving the tanker; but clearly the pronoun can also refer to the tanker itself, as marked by the majority of coders.

(5) 13.1 M: so we have to move the tanker from Corning to Elmira
13.2 : ... uhm
13.3 : but we need an engine for it first

The other pronoun *it* marked with a segment antecedent by multiple coders was in utterance 29.2. It display an ambiguity which is very common in the TRAINS dialogues, between a route and a plan or action of moving trains along this route. More coders chose to mark it as coreferential with *the fastest route* than to give it a segment antecedent.

(6) 28.1 S: the fastest route is via Dansv / is
28.2 : yeah
28.3 : via Dansville
29.1 M: mkay
29.2 : and how long would it take

The observation that personal pronouns are much less likely than demonstratives to refer to abstract objects seems rather robust, in conformance with previous findings (Passonneau, 1993).

Demonstratives were also the easiest markables to identify as having segment antecedents. The eight markables which were given segment antecedents by the most annotators (between 10 and 12 annotators each) were all the pronoun *that*, occurring either as the object of *do/did* (4 instances) or as the subject of *takes/would take* (4 instances); they all referred to plans. The next eight markables, annotated with segment antecedents by 8 or 9 annotators, were also all demonstratives (six *that*, one *those* and one *that way*); they included five references to plans, one which displays the route/plan ambiguity, and two which denote activities that are not plans, for instance the activity of making orange juice.

(7) 21.1 M: um
21.2 : 'bout how long does it take .. to make the oranges into orange juice
22.1 S: that takes an hour

Aside from demonstrative pronouns and the two instances of the pronoun *it* mentioned above, the only markables which reached the criterion of being assigned segment antecedents were definite descriptions with the head nouns *plan* or *way*. The non-demonstrative given segment antecedents by the most annotators was the NP *the plan*, identified by seven annotators. Interestingly enough, some definite descriptions whose form is highly suggestive of a segment antecedent, for example *the plan*, *the current plan* and *the banana problem*, failed to reach the criterion of identification by three annotators.

5 Experiment 2

This experiment tested whether using trained participants and a simplified coding scheme would provide improved results.

5.1 Experimental setup

Materials Dialogue 3.2 from the TRAINS-91 corpus; dialogue 3.1 was used for training.

Participants Four paid undergraduates, all of whom participated in experiment 1.

Procedure Similar to experiment 1, but slightly different marking of segment antecedents as explained above.

5.2 Results

Of the four participants, one didn't mark even a single segment antecedent and was therefore excluded from the study. In order to have more data, we included one of the experimenters as an additional annotator (the experimenter's annotations were produced at the same time as those of the experiment participants and without knowledge of their annotations).

In total, 35 of the 102 markables were identified with a segment antecedent by at least one annotator. Of these, 19 were identified by just one annotator; 15 of those appear to be in error—all and only those marked by one particular annotator, who apparently went for high recall at the expense of precision. The remaining four singular annotations (by three different annotators) appear to be plausible interpretations, so an acceptance criterion that requires agreement by two annotators seems too strong when there are just four annotators in total. We excluded the singular annotations of the overzealous annotator from the analysis.

As in the previous experiment, the annotators appeared to agree overall on the identity of segment antecedents, with a tendency to agree more on the ends of segments than on their beginnings. This is based on an impressionistic evaluation of the data—there are too few data points for a meaningful numerical analysis. This finding holds despite the fact that segment annotation was coarser (that is more constrained) in this experiment.

Also in line with the previous experiment, the most readily identifiable markables were demonstratives—the eight markables assigned segment antecedents by three or four coders were all instances of the pronoun *that*. The six markables which were given segment antecedents by all four annotators clearly referred to plans. The situation is less clear with regard to the two markables which were given segment antecedents by three annotators. The first of those was the word *that* in utterance 10.3, which displays a route/plan ambiguity.

(8) 10.1 S: okay the shortest route would be
10.2 : back through Dansville again
10.3 : that'll take 4 hours
10.4 : and get there
10.5 : get to Corning at 11

Indeed, the remaining annotator marked the word *that* in 10.3 as coreferential with the NP *the shortest route* in 10.1, as did one of the other annotators whose annotation received a discourse antecedent through propagation. A third annotator (the experimenter) marked the word as ambiguous between a segment antecedent and an object antecedent, intending to mark an ambiguity between a plan and a route. Only one annotator marked this unambiguously with a segment antecedent.

The second markable annotated with a segment antecedent by three participants was the word *that* in utterance 13.4. The matter here is more subtle: while the reference of *that* is related to the plan developed in the preceding utterances, it cannot actually denote the plan, but rather a fact about the identity of a plan.

(9) 13.1 M: and when our
[2sec]
13.2 : engine and car .. arrives it at .. Corning
13.3 : I believe we're having it filled with oranges
13.4 : is that correct

Our method of marking antecedents as text regions is not sensitive enough to make such subtle distinctions. The three annotators who chose a segment antecedent for the pronoun *that* in 13.4 marked the preceding utterances (two chose 13.1–13.3 and one chose only 13.3); the fourth marked the pronoun as non-referring.

6 Discussion

If we impose a criterion which requires agreement by at least three of the 20 annotators in experiment 1, we find that 35 of the 181 markables in the dialogue (19.3%) have a plausible interpretation as an anaphor whose antecedent is discussed in preceding discourse but not mentioned by name. A similar figure obtains for experiment 2 after removing the annotations of the participant who appeared to have misunderstood the instructions (20 markables out of 102, or 19.6%). These percentages concur with the 22.6% figure reported by (Eckert and Strube, 2000) for their selection of dialogues from the Switchboard corpus, which is not task-oriented. The figures show that anaphora to entities not mentioned explicitly in the discourse is common enough to warrant treatment.

The fact that many of the segment antecedents in our study turned out to be plans is not surprising, and is due to the dialogues being collected as a planning task. The observation that demonstratives are more likely than other pronouns to have a segment antecedent confirms earlier findings. What is new is the finding that demonstratives are more readily identifiable as elements which require such antecedents—more so than definite descriptions with a highly suggestive head noun. This has implications for writing annotation guidelines, and possibly also for resolution.

The set-based measures show that there is substantial overlap between the annotators regarding the identity of segment antecedents; while far from perfect, this suggests that the task itself is a feasible one, and hopefully can be improved. As for the word-index measures, the fact that a simple model of picking antecedents at a fixed distance from the anaphor accounts for much of what the annotators are doing is in some ways encouraging, as it suggests that the correct vicinity (if not the exact antecedent) could, perhaps, be identified computationally. At the same time, this finding puts in question the added value of human annotation, since annotators have not shown much improve-

ment over the base model. An anonymous reviewer points out that there may be a limit to what we can expect from the annotators because they are in a sense overhearers of the dialogue rather than participants in it, and therefore do not play a part in the grounding process that takes place between the participants (Schober and Clark, 1989). The attainable agreement among annotators may therefore be lower than a reflection of the understanding of dialogue participants.

The same reviewer also suggested out a possible explanation to the fact that annotators agree more on the endings of segment antecedents than on their beginnings, namely that candidate antecedents occur on the right frontier of the discourse structure (Webber, 1991), so their ends tend to coincide. However, the fact that we did not find a difference in chance-corrected agreement between beginnings and ends of antecedents suggests an alternative explanation—perhaps agreement is higher on the ends simply because the space for endings is more compressed. Of course, it could be that both explanations are right and the latter is the result of the former; we would need more experimentation to distinguish between these hypotheses.

It is encouraging that many annotators with little training can converge on roughly similar text regions as antecedents, as it shows that the judgments are not too subjective. Hopefully this should lead to a more systematic study of discourse deixis and discourse antecedents.

Acknowledgments

We wish to thank two anonymous reviewers and the participants of the Essex language and computation reading seminar. This work was in part supported by EPSRC project GR/S76434/01, ARRAU.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1–3):177–210.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of 40th Annual Meeting of the ACL*, pages 80–87, Philadelphia, July.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Derek Gross, James F. Allen, and David R. Traum. 1993. The Trains 91 dialogues. TRAINS Technical Note 92-1, University of Rochester Computer Science Department, July.

Klaus Krippendorff, 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12, pages 129–154. Sage, Beverly Hills, CA.

Klaus Krippendorff, 2004. *Content Analysis: An Introduction to Its Methodology*, chapter 11, pages 211–256. Sage, Thousand Oaks, CA, second edition.

Christoph Müller and Michael Strube. 2003. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 198–207, Sapporo, July.

Costanza Navarretta. 2000. Abstract anaphora resolution in Danish. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue*, Hong Kong, October.

Rebecca J. Passonneau. 1993. Getting and keeping the center of attention. In Madeleine Bates and Ralph M. Weischedel, editors, *Challenges in Natural Language Processing*, pages 179–227. Cambridge University Press, Cambridge.

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of LREC*, volume 4, pages 1503–1506, Lisbon.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Massimo Poesio and Ron Artstein. 2005a. Annotating (anaphoric) ambiguity. In *Proceedings from the Corpus Linguistics Conference Series*, volume 1, Birmingham, England, July.

Massimo Poesio and Ron Artstein. 2005b. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Massimo Poesio, Florence Bruneseaux, and Laurent Romary. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In Marilyn Walker, editor, *Proceedings of the ACL Workshop Towards Standards and Tools for Discourse Tagging*, pages 65–74, College Park, Maryland, June. Association for Computational Linguistics.

Massimo Poesio. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Cambridge, Massachusetts, April-May. Association for Computational Linguistics.

Michael F. Schober and Herbert H. Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.

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