

Scaling Up Anaphora Interpretation

Massimo Poesio

University of Essex

Department of Computer Science

poesio@essex.ac.uk

Abstract

Semantic, discourse, and pragmatic factors such as lexical information, salience, and commonsense inference all play a role in anaphora resolution; yet at least in the case of some anaphoric expressions it is possible to achieve good results using ‘knowledge-poor’ techniques, so that it is already possible to evaluate the performance of algorithms on large numbers of examples. As a result, anaphora resolution techniques are increasingly viewed as potentially useful for real applications. Over the last several years we developed a series of prototypes for definite description resolution that were evaluated on a large scale (thousands of examples). In this paper, I discuss the major bottlenecks we identified, and the solutions adopted so far.

1 Introduction

Anaphora resolution is a good area for pushing the boundaries of research in semantic interpretation. Factors such as lexical semantics, salience, and commonsense inference all play a role in anaphora resolution; yet at least in the case of some anaphoric expressions, such as pronouns, it is possible to achieve good results using ‘knowledge-poor’ techniques, so that it is already possible to evaluate the performance of algorithms on large numbers of examples (Lapin and Leass, 1994; Mitkov, 1998; Tetreault,

2001). So anaphora resolution is increasingly viewed as a potentially useful task, unlike other areas of semantic interpretation such as scope disambiguation, even though the technology is not yet at the point in which it can help practical applications (Morton, 2000).

Over the last several years we have developed a series of prototypes for definite description resolution that were evaluated on a large scale (thousands of examples). In this paper, we discuss the major bottlenecks we identified, and the solutions adopted so far.

2 A Heuristic-Based System for Robust Definite Description Resolution

The starting point for our discussion are the systems for processing definite descriptions (expressions like *the car*) discussed in (Vieira, 1998; Vieira and Poesio, 2000b; Vieira and Poesio, 2000a). An important aspect of this work was that the development of these systems was based on the extensive corpus analysis reported in (Poesio and Vieira, 1998); a second crucial characteristic was that the annotated corpus produced as a result of the analysis was then used to systematically evaluate the systems. We believe both preliminary corpus analysis and extensive evaluation are essential in order to achieve scalability.

The first interesting result of the corpus study discussed in (Poesio and Vieira, 1998) concerned the distribution of definite descriptions. Against our expectations, we discovered that on average, our annotators classified more than half of all definite descriptions as first-mention-

descriptions without an explicit antecedent in the previous discourse. Of these, about 22% were ‘larger-situation’, such as *the Querecho Plains of New Mexico* in (1a); 22% ‘unfamiliar’, like *the economic know-how to steer the city through a possible fiscal crisis* in (1b); and 8% ‘associative’ (or ‘bridging references’ (Hawkins, 1978; Clark, 1977)), like *the kitchen* in (1c).

- (1)
- a. Out here on *the Querecho Plains of New Mexico*, however, the mood is more upbeat trucks rumble along the dusty roads and burly men in hard hats sweat and swear through the afternoon sun.
 - b. They wonder whether he has *the economic know-how to steer the city through a possible fiscal crisis*, and they wonder who will be advising him.
 - c. Once inside, she spends nearly four hours measuring and diagramming each room in *the 80-year-old house*, gathering enough information to estimate what it would cost to rebuild it. While she works inside, a tenant returns with several friends to collect furniture and clothing. One of the friends sweeps broken dishes and shattered glass from a countertop and starts to pack what can be salvaged from *the kitchen*.

Another important finding of the corpus analysis was that there was little agreement among our annotators (61% total), except concerning whether a definite description was first mention or subsequent mention ($\kappa = .73$). Disagreement on identifying bridging references was particularly bad ($\kappa = .24$), but it was also difficult to get the annotators to agree on which entities could be assumed to be known as part of ‘general knowledge’.

The result concerning the prevalence of discourse-new descriptions led us to develop heuristic methods for identifying them, based in part on the discussion in (Hawkins, 1978). Particularly effective heuristics were checking whether the head of the clause was a predicate like *fact*, *result*; and if the definite was post-modified by a relative clause. Overall, these heuristics achieved a recall R=69% and preci-

sion P=72%.

For anaphora, we developed heuristics for dealing with premodification, to avoid, e.g., a match between *the red car* and *the blue car*, while allowing *the house* to match *the 80-year-old Victorian house*. We also develop segmentation heuristics, to restrict the number of potential antecedents. Overall, the system achieved 62% recall and 83% precision on anaphoric same head descriptions.

Finally, our work on bridging descriptions was based on a classification of bridging descriptions into classes depending on the sort of information needed to resolve them (Poesio et al., 1997; Vieira and Teufel, 1997; Vieira, 1998) which heavily relied on previous classifications proposed by (Clark, 1977; Sidner, 1979; Strand, 1996). We identified the following classes:

- cases based on lexical relations, such as synonymy, hypernymy and meronymy, that can be found in a lexical database such as WordNet (Fellbaum, 1998)—as in *the flat ... the living room*;
- bridging descriptions in which the antecedent is a proper name and the description a common noun, whose resolution requires some way of recognizing the type of object denoted by the proper name (as in *Bach ... the composer*);
- cases in which the anchor is not the head noun but a noun modifying an antecedent, as in *the company has been selling discount packages ... the discounts*
- cases in which the antecedent (anchor) is not introduced by an NP but by a VP, as in *Kadane oil is currently drilling two oil wells. The activity ...*
- descriptions whose antecedent is not explicitly mentioned in the text, but is implicitly available because it is a discourse topic—e.g., *the industry* in a text referring to oil companies;
- cases in which the relation with the anchor is based on more general commonsense

knowledge, e.g., about cause-consequence relations.

Our algorithms for bridging descriptions were based on a combination of access to WordNet 1.6 and heuristics (Poesio et al., 1997; Vieira and Teufel, 1997; Vieira and Poesio, 2000a). The system achieved R=29% and P=38% on bridging descriptions.

Two versions of the system were evaluated. One version only resolved directly anaphoric definites and identified discourse-new descriptions; this system had R=53%, P=76%. Version II also attempted to find the anchors of bridging descriptions; this version had a higher recall (57%) but a lower precision (70%).

3 Problems with the early system

The results just mentioned, while comparable to those obtained by other systems on the same task, are clearly not sufficient for real applications; our efforts since then have concentrated on overcoming these problems. Our evaluation of the system pointed out a number of areas where improvement was needed.

A fundamental issue was the need to find better annotation techniques, which would overcome the disagreement problems observed in the first study, so as to improve our evaluation methods.

As for the algorithms proper, the worse results by far were obtained on bridging descriptions. Part of the problem was that our system didn't include methods for resolving bridging descriptions based on 'causal' information or on thematic roles, but even on the 19% of bridging descriptions depending for their resolution on information contained in WordNet (synonymy, hyponymy, and meronymy) we only had 39% recall. Better lexical resources were needed for this purpose, but improving on these results was clearly going to be difficult. One type of bridging descriptions in which we could expect to improve the performance of the system was definite descriptions referring to antecedents introduced by proper names. Our methods for resolving these bridging descriptions worked better than those for other classes

(R=66%, P=95%) but state-of-the-art named entity recognition systems can achieve about 90% precision and recall (Mikheev et al., 1999). More in general, we needed to improve our evaluation methods.

The second main problem was the system's lack of encyclopedic knowledge, needed to handle 'larger situation' references. This was the main source of problems for the 'discourse new' class (66% precision, as opposed to 76% for the 'unfamiliar' cases).

On the issue of salience, our early work (Vieira and Poesio, 2000b) had indicated that not using any segmentation heuristic at all led to more than one possible interpretation for about 10% of anaphoric expressions. Our best segmentation heuristics (4 sentences window with recency) achieved a recall of 75.96% and a precision of 87.77%. On the other hand, our system did not include methods for tracking the local focus, in part because we were aware that unless a very good focus tracking mechanism was available, its inclusion could lead to worse results (Azzam et al., 1998), in part because we didn't expect a focus tracking mechanism to be essential for definite description resolution (unlike, say, pronoun resolution). Subsequent work however indicated that this expectation was incorrect (see below).

In the last few years, we have tried to overcome some of these problems. Our work included the development of better annotation methods, and of a more usable corpus; additional corpus-based studies of bridging references and the impact of salience on anaphora; methods for acquiring lexical knowledge; better integration with parsing technology and work on named entity recognition. We discuss these developments in turn.

4 Additional Empirical Studies and Further Corpus Annotation

The early work did not result in a corpus annotated in a standardized way that could be easily used by other groups. Also, the scheme we used wasn't completely satisfactory. For one thing, because of the characteristics of our

domain, we had not included a class covering visual references, although this would be easy to add. Secondly, and most importantly, our method for annotating bridging references didn't guarantee agreement. Third, although our work had revealed that many definite NPs could simultaneously belong to two classes - e.g., be directly anaphoric on one entity, while bridging on another one - our scheme wouldn't allow our annotators to mark a definite NP as belonging to more than one class. Finally, we didn't have a method for marking up genuinely ambiguous cases. Because we felt that a better annotation technology was a prerequisite to better evaluation methods, hence to the development of better systems, we devoted quite a lot of attention to this.

The MATE scheme

In the MATE project we developed an XML-based annotation scheme (Poesio et al., 1999) integrating the results from (Poesio and Vieira, 1998) with ideas from Passonneau (Passonneau, 1997), MUC-7 (Chinchor and Sundheim, 1995) and (Bruneseaux and Romary, 1997). In the MATE scheme NPs are not classified in different classes; instead, both direct and indirect anaphoric links are marked, as in the MUC-7 scheme, and the classification of a NP depends on whether it is related to another entity by one such link. However, the scheme also fundamentally differs from MUC-7 scheme, in that the information about NPs is split among two XML elements. Each NP in the text is tagged with an `<ne>` tag, as follows:

```
(2) <ne ID="ne07" ... >
    Scottish-born, Canadian based jeweller,
    Alison Bailey-Smith</ne>
    ...
    <ne ID="ne08"> <ne ID="ne09">Her</ne>
    materials</ne>
```

Anaphoric relations are then annotated by means of a separate `<ante>` element specifying relations between `<ne>`s.¹ An `<ante>` element includes one or more `<anchor>` element,

¹The `<ante>` element derives from the `LINK` element in the annotation scheme proposed by the Text Encoding Initiative and used by Bruneseaux and Romary, 1997).

one for each plausible antecedent of the current discourse entity. E.g., the anaphoric relation in (2) between the possessive pronoun with ID = "ne09" and the proper name with ID = "ne07" is marked as follows:

```
(3) <ante current="ne09">
    <anchor ID="ne07" rel="ident" ... >
    </ante>
```

Separating the `<ante>` element from the `<ne>` element makes it possible to mark an NP being both directly anaphoric to one entity, and bridging on a second one, if necessary: this can be done by specifying two separate `<ante>` elements. It is also possible to mark ambiguous anaphoric expressions, by specifying more than one `<anchor>` element. References to visually accessible entities not previously mentioned in the discourse can be marked by including in the annotated file special `<universe>` elements, one for each such entity, and by specifying these elements as anchors for an anaphor (this idea is reminiscent of the technique used for landmark references in the MAPTASK corpus).

The GNOME scheme

The MATE scheme was further developed and tested in the GNOME project, concerned with generation of noun phrases (Poesio, 2000a). The project produced a preliminary version of a corpus annotated according to a scheme described in (Poesio, 2000b) and containing texts from several domains, including descriptions of museum objects in the domain of the ILEX and SOLE projects (Oberlander et al., 1998), pharmaceutical leaflets in the domain of the ICONOCLAST project (Scott et al., 1998), and tutorial dialogues from the Sherlock corpus collected at the University of Pittsburgh (Lesgold et al., 1992; Di Eugenio et al., 1997). Each sub-corpus contains about 6,000 NPs.

While the annotation of the GNOME corpus is not yet entirely complete, subsets of the corpus have been used for studying local focus, including its relationship with bridging (Poesio et al., 2000; Poesio et al., 2002) and the correlation between definiteness and functionality proposed by Loebner, 1987) (Poesio, 2001) (see below). This work revealed a few problems with

the MATE scheme, which led us to backtrack on some of the changes discussed above:

1. The scheme for marking up visual deixis proved too complicated for cases in which the annotator had to deal with real objects, rather than abstract characterizations such as the MAPTASK maps with few landmarks, all clearly identified. The additional complexity of introducing universe elements for all the objects referred to is only justified if the goal of the annotation is to evaluate a system for resolving such references; if the only purpose is to classify a noun phrases as visually referring or not, a boolean attribute `<visual-deixis>` is much easier, and can be marked reliably provided that we don't ask the annotators to distinguish between 'visible' and 'immediate' deixis in the sense of Hawkins, 1978).
2. This problem occurs in spades with discourse deixis, where identifying the antecedents in the text may in general prove too difficult (Eckert and Strube, 2001; Poesio and Reyle, 2001). Again, using a boolean attribute to mark a NP as discourse deictic is much simpler to do, and can be done reliably (Modjeska-Nygren and Poesio, in preparation).

Finally, we will note that the MATE / GNOME scheme does not include any methods for further classifying first mention definites into 'larger situation' and 'unfamiliar' definites. The difficulty here is not from the annotation scheme point of view—doing this would simply involve adding another boolean attribute— but in finding instructions that would allow the annotators to do the job reliably. We expect to try this in the future.

We expect the GNOME corpus will allow us a better evaluation of our system's performance on bridging descriptions. We also hope to eventually make this corpus available so as to facilitate the comparison between anaphora resolution systems.

5 Acquiring Lexical Knowledge

Pure vector-based methods

Some researchers's solution to the shortcomings of WordNet has been to augment it by adding further information (Harabagiu, 1998). We decided instead to test vector-based methods for unsupervised lexical acquisition (Lund et al., 1995; Schütze, 1997). These methods are based on the assumption that the meaning of each (sense of a) word w is simply a vector of 'features'—which, in the simplest cases, are simply other words that occur in the vicinity of w . The reason for our interest in these methods were the results by Lund et al., 1995), who found a high correlation between the lexical associations acquired in this way and the lexical associations discovered by means of semantic priming experiments (Moss et al., 1995). Lund et al's results encouraged us to test whether the anchor for a bridging description could be found simply by finding the antecedent most strongly associated with that description—henceforth, the PRIMING HYPOTHESIS.²

In a series of experiments discussed in (Schulte im Walde, 1997; Poesio et al., 1998), we used the BNC corpus to acquire this type of lexical meanings for the bridging descriptions and their antecedents in the cases of bridging references tested in (Vieira and Poesio, 2000a). Part of our goal was to find the best values for the corpus acquisition method. We tried various window sizes, vicinity measures, and various types of corpus preprocessing, including lemmatization and tagging; the best results were obtained with the configuration using windows of size 10, lemmatized but untagged, and using the cosine metric as a vicinity measure. With this configuration, we found that the priming hypothesis didn't hold in the simple form sketched above—for only 29% of bridging descriptions the strongest lexical associate in the previous five sentences was also the correct anchor. Our subsequent case analysis suggested however that while the results were not very good, the problems had only in part to

²This possibility has been raised, among others, by Carter, 1987).

do with the lexical knowledge we had acquired.

We classified each suggested resolution as either Acceptable, F (the resolution is arguably the closest semantic associate of the bridging description, but the 'correct' anchor is more in focus), Lexically plausible (although the desired one should have perhaps be classified as closer) and Wrong. 29% of resolutions were Acceptable, 21.2% F, 9.8% in class L, and 39.9% W. For at least certain types of lexical association, the results with the automatically acquired lexica were comparable to those we had obtained with WordNet: in particular, the accuracy for synonymy was 36%, identical to that obtained with WordNet. The worse results were obtained with meronymy (accuracy =16.7% vs 25% with WordNet) and hyponymy (accuracy = 14.3%, vs. 57.1%) (Schulte im Walde, 1997; Poesio et al., 1998). And in all of these cases, we found that we were far from achieving the best possible performance, at least for synonymy: for example, doubling the size of the training corpus from 50M to 100M words increased the accuracy by almost 50%.

Part of the problem was the resolution method, and in particular the fact that we didn't keep track of the current local focus. The 'window' heuristic for tracking global salience was simultaneously too restrictive and not restrictive enough. On the one hand, fully 19.6% of the actual anchors were outside the 5-sentence window we were using. On the other hand, in a number of cases the anchor suggested by the algorithm is can be argued to be semantically closer than the actual antecedent, which is however the local focus or at least closer. In one case, the algorithm suggested the lexical associate *customer* as the antecedent of *market*, whereas the actual anchor is the (genre-specific) hyponym *phone service*. In an extreme case, the algorithm picks up *investigative companies* as antecedent for *the company*, whereas the actual antecedent is a specific company, *Pinkerton*.

These results suggested, first, the need to integrate bridging resolution with a focus tracking mechanism; second, that at least for synonymy, these automatically acquired resources

were comparable to WordNet, and increasing the accuracy might just be a matter of increasing the training corpus. On the other hand, we felt that these methods didn't work too well for hyponymy and meronymy. These considerations led us, on the one hand, to undertake extensive empirical investigations of local and global salience, discussed below. In order to improve our lexical resources, we considered using a mixture of lexical sources for the different types of bridging descriptions, acquired in different ways.

Syntactic patterns and the acquisition of meronymic information

We were particularly interested in testing whether using syntactic information would help, given the results by (Grefenstette, 1993) and especially Hearst's work on automatically acquiring hyponymy information (Hearst, 1998). Concentrating on the meronymy case,³ we hypothesized that we would get better information about meronymy by taking as 'mereological neighbors' of a word *W* not all words occurring in its vicinity, but only those occurring in certain syntactic constructions, such as *the Z of W* or *Z's W*. E.g., *window* would be considered a 'mereological neighbor' or *car* if it frequently occurred in constructions such as *the car's window*. Again, we used the British National Corpus to train our lexical bases, and we used the same texts as (Vieira and Poesio, 2000a; Poesio et al., 1998) to test our models. The algorithms used point-wise mutual information (Brown et al., 1992) to identify the closest 'mereological antecedent' of each bridging description. The results were very promising: we obtained 66.7% recall and 72.7% precision on the the mereological descriptions, as opposed to 25% using WordNet and 16.7% using pure vector association obtained before (Poesio et al., 2002). This suggests that the best results are going to be obtained by employing a combination of specialized lexical acquisition methods rather than by a single method.

³The acquisition of information about hyponymy has been intensively studied in the last years—see, e.g., (Caraballo, 1999).

6 Local Salience and Bridging References

As said above, our initial system did not include focus-tracking algorithms, as they didn't appear to be essential for definite description resolution (as opposed to pronoun resolution), and work such as (Azzam et al., 1998) seemed to suggest that adding focus tracking need not improve the performance of a system. Our later results about bridging, however, led us to reconsider this position. We have undertaken an extensive study of the claims of the best-known theory of salience, Centering Theory (Grosz et al., 1995; Walker et al., 1998), to investigate its impact on pronominalization and more in general on anaphora (Poesio et al., 2000).

Our results suggest that the connection between Centering's notion of 'focus' –the BACKWARD-LOOKING CENTER, or CB–and bridging references is complex. On the one hand, bridging references do play a central role in maintaining the local focus / CB: if only direct anaphora is allowed to realize entities in an utterance, some of the central claims of Centering, namely Constraint 1 (stating that each utterance has exactly one CB) do not hold. This is illustrated by examples like (4), in which utterance u1 is followed by four utterances. Only the last of these directly refers to the set of egg vases introduced in u1, while they all contain implicit references to these objects. In (4a), (entity) coherence is maintained by the bridging reference (*the furniture*) rather than by direct reference. (See below.)

- (4) (u1) These “egg vases” are of exceptional quality: (u2) basketwork bases support egg-shaped bodies (u3) and bundles of straw form the handles, (u4) while small eggs resting in straw nests serve as the finial for each lid. (u5) Each vase is decorated with inlaid decoration: ...

On the other hand, not all bridging references refer to the CB: in our corpus, about 190 bridging references have the CB as their anchor,

whereas 372 haven't.

7 Topics for Future Research and Conclusions

The main claims made in this paper are that anaphora resolution, especially for 'fuller' noun phrases such as definite descriptions, is an excellent way of investigating semantic issues; and that developing scalable systems involves being able to evaluate them, e.g., by comparing their output with an annotated corpus.

We found that resolving bridging references is the hardest problem - in part because 'knowledge poor' methods don't work very well in this case, in part because it's not clear how humans themselves deal with them. We found however that the agreement problem is lessened by concentrating on a smaller range of cases.

Long term, we need robust methods for focus tracking, and ways for acquiring encyclopedic and causal knowledge.

References

- Azzam, S., K. Humphreys, and R. Gaizauskas, 1998. Evaluating a focus-based approach to anaphora resolution. In *Proc. of the 36th ACL and 17th COLING*. Montreal, CA: ACL.
- Brown, P., V. J. D. Della Pietra, P. V. DeSouza, J. C. Lai, and R. L. Mercer, 1992. Class-based n-grams models of natural language. *Computational Linguistics*, 18(4):467–479.
- Bruneseaux, F. and L. Romary, 1997. Codage des références et coréférences dans le dialogues homme-machine. In *Proc. of ACH-ALLC*. Kingston.
- Caraballo, S. A., 1999. Automatic acquisition of a hypernym-labeled noun hierarchy from text. In *Proc. of the ACL*.
- Carter, D. M., 1987. *Interpreting Anaphors in Natural Language Texts*. Chichester, UK: Ellis Horwood.
- Chinchor, N. A. and B. Sundheim, 1995. Message Understanding Conference (MUC) tests of discourse processing. In *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*. Stanford.

- Clark, H. H., 1977. Bridging. In P. N. Johnson-Laird and P.C. Wason (eds.), *Thinking: Readings in Cognitive Science*. London and New York: Cambridge University Press.
- Cristea, D., N. Ide, D. Marcu, and V. Tablan, 2000. Discourse structure and co-reference: An empirical study. In *Proc. of COLING*. Saarbruecken.
- Di Eugenio, B., J. D. Moore, and M. Paolucci, 1997. Learning features that predict cue usage. In *Proc. of the 35th ACL*. Madrid.
- Eckert, M. and Strube, M., 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*.
- Fellbaum, Christiane, editor. 1998. *WordNet: An electronic lexical database*. The MIT Press.
- Fox, B. A., 1987. *Discourse Structure and Anaphora*. Cambridge, UK: Cambridge University Press.
- Grefenstette, G., 1993. SEXTANT: extracting semantics from raw text. *Heuristics*.
- Grosz, B. J., A. K. Joshi, and S. Weinstein, 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202–225. (The paper originally appeared as an unpublished manuscript in 1986.).
- Grosz, B. J. and C. L. Sidner, 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- Harabagiu, S. 1998. *WordNet-based inference of textual context, cohesion and coherence*. PhD Dissertation, University of Southern California.
- Hawkins, J. A., 1978. *Definiteness and Indefiniteness*. London: Croom Helm.
- Hearst, M. A., 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.
- Hearst, M. A., 1998. Automated discovery of wordnet relations. In C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*. MIT Press.
- Lappin, S. and H. J. Leass, 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562.
- Lesgold, A.M., S.P. Lajoie, M. Bunzo, and G. Eggan, 1992. SHERLOCK: A coached practice environment for an electronics troubleshooting job. In J. Larkin and R. Chabay (eds.), *Computer assisted instruction and intelligent tutoring systems: Shared issues and complementary approaches*. Hillsdale, NJ: Erlbaum, pages 201–238.
- Loebner, S., 1987. Definites. *Journal of Semantics*, 4:279–326.
- Lund, K., C. Burgess, and R. A. Atchley, 1995. Semantic and associative priming in high-dimensional semantic space. In *Proc. of the 17th Annual Conference of the Cognitive Science Society*.
- Mikheev, A., M. Moens, and C. Grover, 1999. Named Entity recognition without gazetteers. In *Proc. of EACL*. Bergen, Norway: EACL.
- Mitkov, R., 1998. Robust pronoun resolution with limited knowledge. In *Proc. of the 18th COLING*. Montreal.
- Morton, T., 2000. Coreference for NLP applications. In *Proc. of the 38th ACL*. Hong Kong.
- Moss, H. E., R. K. Ostrin, L. K. Tyler, and W. D. Marslen-Wilson, 1995. Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*:863–883.
- Oberlander, J., M. O'Donnell, A. Knott, and C. Mellish, 1998. Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Hypermedia and Multimedia*, 4:11–21.
- Passonneau, R., 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript.
- Poesio, M., 2000a. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proc. of the 2nd LREC*. Athens.
- Poesio, M., 2000b. *The GNOME Annotation Scheme Manual*. University of Edinburgh, HCRC and Informatics, Scotland, fourth version edition. Available from <http://www.hcrc.ed.ac.uk/~gnome>.
- Poesio, M., 2001. Definiteness: Familiarity or functionality? an empirical investigation. Paper presented at Sinn und Bedeutung VI.
- Poesio, M., F. Bruneseaux, and L. Romary, 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In M. Walker (ed.), *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging*.
- Poesio, M., H. Cheng, B. Di Eugenio, J. M. Hitzenman, and R. Stevenson, 2002. A corpus-based evaluation of centering theory. Submitted.

- Poesio, M., H. Cheng, R. Henschel, J. M. Hitzeman, R. Kibble, and R. Stevenson, 2000. Specifying the parameters of Centering Theory: a corpus-based evaluation using text from application-oriented domains. In *Proc. of the 38th ACL*. Hong Kong.
- Poesio, M., T. Ishikawa, S. Schulte im Walde, and R. Vieira, 2002. Acquiring lexical knowledge for anaphora resolution. In *Proc. of the 3rd LREC*. Las Palmas.
- Poesio, M. and U. Reyle, 2001. Underspecification in anaphoric reference. In E. Thijsse H. Bunt, I. van der Sluis (ed.), *Proc. of the Fourth International Workshop on Computational Semantics*. Tilburg.
- Poesio, M., S. Schulte im Walde, and C. Brew, 1998. Lexical clustering and definite description interpretation. In *Proc. of the AAAI Spring Symposium on Learning for Discourse*. Stanford, CA: AAAI.
- Poesio, M. and R. Vieira, 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216. Also available as Research Paper CCS-RP-71, Centre for Cognitive Science, University of Edinburgh.
- Poesio, M., R. Vieira, and S. Teufel, 1997. Resolving bridging references in unrestricted text. In R. Mitkov (ed.), *Proc. of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*. Madrid. Also available as HCRC Research Paper HCRC/RP-87, University of Edinburgh.
- Reichman, R., 1985. *Getting Computers to Talk Like You and Me*. Cambridge, MA: The MIT Press.
- Schulte im Walde, S., 1997. Resolving bridging descriptions in high-dimensional space. Studienarbeit, Universities of Stuttgart and Edinburgh.
- Schütze, H., 1997. *Ambiguity Resolution in Language Learning*. Stanford: CSLI.
- Scott, D., R. Power, and R. Evans, 1998. Generation as a solution to its own problem. In *Proc. of the 9th International Workshop on Natural Language Generation*. Niagara-on-the-Lake, CA.
- Sidner, C. L., 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.
- Strand, K., 1996. A taxonomy of linking relations. Manuscript. A preliminary version presented at the Workshop on Indirect Anaphora, Lancaster University, 1996.
- Tetreault, J. R., 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.
- Vieira, R., 1998. *Definite Description Resolution in Unrestricted Texts*. Ph.D. thesis, University of Edinburgh, Centre for Cognitive Science.
- Vieira, R. and M. Poesio, 2000a. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4).
- Vieira, R. and M. Poesio, 2000b. Processing definite descriptions in corpora. In S. Botley and T. McEnery (eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*, chapter 10. Amsterdam / New York: John Benjamins, pages 189–212.
- Vieira, R. and S. Teufel, 1997. Towards resolution of bridging descriptions. In *Proc. of the 35th Joint Meeting of the Association for Computational Linguistics*. Madrid.
- Walker, M. A., A. K. Joshi, and E. F. Prince (eds.), 1998. *Centering Theory in Discourse*. Clarendon Press / Oxford.