ARQAS - Venice, 23-24 June 2003.

An Empirical Evaluation of Pronoun Resolution and Clausal Structure

Joel Tetreault
Department of Computer Science
University of Rochester
Rochester, NY, 14627, USA
tetreaul@cs.rochester.edu

James Allen
Department of Computer Science
University of Rochester
Rochester, NY, 14627, USA
james@cs.rochester.edu

Abstract

This paper presents an automated empirical evaluation of the relationship between clausal structure and pronominal reference. Past work has theorized that incorporating discourse structure can constrain the search space in the resolution of pronouns since discourse segments, and thus potential antecedents, can be made inaccessible as the discourse progresses and the focus changes. However, very little empirical work has been done to evaluate these claims. In this study, we develop an automated system and use a corpus annotated for RST and coreference to test whether basic formulations of these claims hold. In particular, we develop and evaluate two pronoun resolution algorithms that incorporate clausal and discourse structure. The first is based on Grosz and Sidner's theory of discourse structure and the second is based on Cristea et al.'s Veins Theory. Our results show that incorporating basic clausal structure does not improve performance.

1 Introduction

In this paper we present an automated corpus-based analysis using Rhetorical Structure Theory (Mann and Thomson, 1988) to aid in pronoun resolution. Most implemented pronoun resolution methods in the past have used a combination of focusing metrics, syntax, and light semantics[1], but very few have incorporated discourse information or clausal segmentation. It has been suggested that discourse structure can improve the accuracy of reference resolution by closing off unrelated segments of discourse from consideration. However, until now, it has been extremely difficult to test this theory because of the difficulty in annotating discourse structure and relations reliably and for a large enough corpus. What limited empirical work that has been done in this area has focused primarily on how structure can constrain the search space for antecedents (Poesio and Di Eugenio, 2001; Ide and Cristea, 2000) and their results show that it can be effective. In this paper, we use a different metric, simply, how many pronouns one can resolve correctly with a constrained search space.

This paper builds on preliminary research discussed in (Tetreault, 2002) in which the RST-tagged Treebank (Carlson et al., 2001) corpus of Wall Street Journal articles merged with coreference information is constructed to provide a testing ground for the claims above. In addition, an existing pronoun resolution system (Byron and Tetreault, 1999) is augmented with modules for incorporating the information from the corpus: discourse structure and relations between clauses. With this testbed system, we evaluate two algorithms based on leading theories of decomposing discourse: Grosz and Sidner (1986) and Veins Theory (Cristea et al., 1998). Our results show that basic methods of decomposing discourse do not improve performance of pronoun resolution

[1] See Mitkov (2000) for a leading method.

methods.

In the following section we discuss theories that relate discourse and anaphora. Next we discuss two evaluations: the first determines a baseline algorithm to be compared against and the second tests the two new algorithms using RST. Finally, we close with results and discussion.

2 Background

2.1 Discourse Structure

We follow Grosz and Sidner's (1986) work in discourse structure in implementing some of our clausal-based algorithms. They claim that discourse structure is composed of three interrelated units: a linguistic structure, an intentional structure, and an attentional structure. The linguistic structure consists of the structure of the discourse segments and an embedding relationship that holds between them.

The intentional component determines the structure of the discourse. When people communicate, they have certain intentions in mind and thus each utterance has a certain purpose to convey an intention or support an intention. Grosz and Sidner (henceforth G&S) call these purposes "Discourse Segment Purposes" or DSP's. Given the nesting of DSP's, the intentional structure forms a tree, with the root of the tree being the main intention of the discourse. The intentional structure is more difficult to compute since it requires recognizing the discourse purpose and the relation between intentions.

The final structure is the attentional state, which is responsible for tracking the participant's mental model of what entities are salient or not in the discourse. It is modeled by a stack of focus spaces which is modified by changes in the intentional state. The set of focus spaces available at any time is the focusing structure. Focus spaces are removed (popped) and added (pushed) from the stack depending on their respective discourse segment purpose and whether or not their segment is opened or closed. The key points about attentional state are that it maintains a list of the salient entities, prevents illegal access to blocked entities, is dynamic, and is dependent on the intentional state.

To our knowledge, there has been no large-scale annotation of corpora for intentional structure. In our study, we use RST (Mann and Thompson, 1988) to approximate the intentional structure in Grosz and Sidner's model. With some sort of segmentation and a notion of clauses one can test pushing and popping, using the depth of the clause in relation to the surrounding clauses.

Using RST to model G&S discourse structure is not without precedent. Moser and Moore (1996) first claimed that the two were quite similar in that both had hierarchal tree structures and that while RST had explicit nucleus and satellite labels for relation pairs, DSP's also had implicit salience labels, calling the primary sentence in a DSP a "core," and subordinating constituents "contributors." However, Poesio and DiEugenio (2001) point out that an exact mapping is not an easy task as RST relations are a collection of intentional but also informational relations and it is not clear how to handle subordinating DSP's of differing relations to model pushes and pops of the attentional stack.

2.2 Veins Theory

Veins Theory (Cristea et al., 1998; Ide et al., 2000) is an extension of Centering Theory from local to global discourse. The empirically tested method makes use of discourse structure (RST trees) to determine the accessibility of referents. The theory assumes that only a subset of the clauses preceding the anaphor are actually relevant to successfully interpreting the anaphor. This subset (domain of referential accessibility, or DRA) is determined by the interaction of the tree hierarchy and whether a clause is a nucleus or a satellite. As a result of this pruning effect, the theory has the advantage over knowledge-poor approaches to pronoun resolution since it constrains the search space for a pronoun.

Using RST as the basis for their discourse representation, terminal nodes in the binary tree represent the clauses of the discourse, and non-terminal nodes represent the rhetorical relations. The DRA for a clause is computed in two steps. First, the "head" of each node is computed bottom-up by assigning a number to each terminal node. Non-terminal nodes are labeled by taking the union of the heads of its nuclear children. The second step, computing the "vein," is a top-down method. First, the vein of the root of the tree is the head. For every nuclear node, if it has a left sibling that is a satellite, its vein is the union of the head of the child and its parent vein, otherwise it inherits its parent's vein only. For every satellite node, if the node is the left child of its parent then its vein is the union of its head with the parent's vein. Otherwise, its vein is the union of its head with the parent's vein but with all prior left satellites removed. Finally, the DRA for a clause is simply all the nodes in the clause's vein that precede it. Intuitively, if a node has parents that are all nuclei, it will be more accessible to other entities since it is highly salient according to Veins Theory (VT). However, satellites serve to restrain accessibility.

3 Baseline Selection

Determining the usefulness of incorporating discourse information in reference resolution requires a large corpus annotated with coreference and clausal information, and a system to try different algorithms. In the following sections we discuss our corpus, our testbed system for extracting noun-phrase entities, and finally the algorithms and their results. After testing each algorithm on the same corpus, the best one would be selected as the baseline algorithm. If discourse or clausal information is used correctly we should see an improvement over the baseline algorithm.

3.1 Corpus Description

The test corpus was constructed by merging two different annotations of a subset of the Penn Treebank (Marcus et al., 1993). The news articles cover such varied topics as reviews of TV shows and the Japanese economy. The portion of the Treebank consists of 52 Wall Street Journal articles which includes 1241 sentences and 454 non-quoted third person pronouns that refer to noun-phrase entities. 10 of the pronouns have long-distance antecedents, where the antecedent is found two or more sentences away from the pronoun.

Carlson et al. (2001) annotated those articles with rhetorical structure information in the manner of Mann and Thompson (1988) with very high annotator reliability. This annotation breaks up each discourse into clauses connected by rhetorical relations. So from this work there is a decomposition of sentences into a smaller units (a total of 2755 clauses) as well as a discourse hierarchy for each article and relations between pairs of segments. The corresponding Penn Treebank syntactic structures for each sentence were also annotated with coreference information in the same manner as Ge et al. (1998). This meant that all third-person pronouns were marked with a specific identification number and all instances of the pronoun's antecedent were also marked with the same id. In addition, the Penn Treebank includes annotations for the syntactic tree structures of each sentence so syntactic attributes such as part-of-speech and number information were extracted. Also, each noun phrase entity was marked manually for gender information.

Finally, the RST corpus and the Penn Treebank coreference corpus were merged such that each discourse entity (in this case, only noun-phrases) had information about its syntactic status, gender, number, coreference, etc. and the following discourse information: the clause it is in, the depth of the clause in the RST tree, and the rhetorical relations that dominate the clause. The Penn Treebank data and only the clausal breakdown of each sentence are used in this evaluation. In the second evaluation, all of the RST data comes into play.

3.2 Algorithms

One of the problems with reporting the performance of a pronoun resolution algorithm is that researchers often test on different corpora so it is hard to compare results. For example, an algorithm tested on a news corpus may perform differently on a corpus of short stories. In this particular experiment, we have a common corpus to test different algorithms, with the goal of simply selecting the best one to use as a baseline for comparison with schemes that incorporate clausal information. We examine three well-known pronoun resolution methods: Left-Right Centering (Tetreault, 1999), the S-list algorithm (Strube, 1998), and Brennan et al.'s centering algorithm (1987), in addition to a naive metric. The naive metric involves searching through a history list starting with the last mentioned item and selecting the first one that meets gender, number, and syntactic constraints. All four algorithms are primarily syntax-based. Because of this limitation they should not be expected to fare too well in interpreting pronouns correctly since proper interpretation requires not only syntactic information but also semantics and discourse information.

Each algorithm was tested on the corpus in two different versions (see Figure 1): the first is the conventional manner of treating sentences as the smallest discourse unit (S); the second involves splitting each sentence into clauses specified by the RST annotations (C).

The (S) results agree with the larger study of the same algorithms in Tetreault (2001) - that the LRC performs better than the other two algorithms and that on a new corpus, one would expect the algorithm to resolve 80% of the pronouns correctly.

The (C) results are a first stab at the problem of how to incorporate clausal structure into pronoun resolution algorithms. The result is negative since each algorithm has a performance drop of at least 3%. The main result for our purposes is that LRC performs the best and thus is selected a the baseline algorithm.

4 Algorithms

In this section, we describe two pronoun resolution algorithms that use clausal structure to constrain the search for antecedents. We also describe a series of corpus transformations that each algorithm is tested on.

4.1 Grosz and Sidner Stack Approximation Algorithm

Based on Grosz and Sidner's pushing and popping discourse structure, we work under the simple assumption that an entity is inaccessible if it is more embedded in the RST tree than the referring entity, meaning if we were explicitly tracking the attentional state, that embedded utterance would have been popped before our current utterance was processed.

Thus the Grosz and Sidner approximation (henceforth G&S) works only by considering the depth of past clauses. The algorithm is as follows: for each pronoun the attentional stack is constructed on the fly since we have perfect information on the structure of the discourse. The search works by looking through past clauses that are either at the same depth or closer to the root than the previous clause visited. The reasoning is that embedded segments that are farther from the root are not related to the entities that follow them. If they were, they would share the same embedding. Clauses at the same depth can be viewed as being in the same discourse segment and clauses that are closer to the root can be viewed as dominating the current clause. In addition, the previous clause is always searched even if it is a lower depth. This follows Walker's (2000) analysis which found that reference can occur between two utterances even if they are split by a segment boundary.

Figure 2 shows how this works. Assume that clauses closest to the left are the closest to the root of the tree (lower depth). When searching for an antecedent for a pronoun in C7, first search all preceding entities in C7, if one is not found, then go back clause by clause until one is found. So the search order would be C6 (since the previous utterance is automatically search regardless of depth) then skip over C5 since it is more embedded than the current clause C6. C4, however, is accessible since it is the same depth as C4.

4.2 Veins Algorithm

The original formulation off Veins Theory is a metric of accessibility not resolution since it does not specify how to search the DRA or how to search clauses within the DRA. The algorithm presented here uses the constraints of VT within the framework of LRC. The algorithm is as follows: for every pronoun, search the clauses of its DRA from most recent (the current clause) to least recent, from left-to-right. If an antecedent is not found within the DRA, the LRC algorithm is used to find a suitable antecedent by searching all past clauses. This approximation accords with the VT claim that referents outside the DRA incur a higher processing load on the interpreter. This "backup mechanism" results in a 14% boost in performance.

In terms of long-distance pronominalization, the original Veins formulation was unable to resolve 6 of the 10 cases when treating sentences as the minimal discourse unit, and when considering clauses, was unable to resolve 9 of the 10 cases. All of these were pronouns and antecedents in attribution relations.

4.3 Corpus Transformations

Tetreault (2002) showed that using the RST tree in the Grosz and Sidner approach produced very poor results (in the 50% range). We believe that the RST decomposition produces too fine a segmentation and

Algorithm	Right (S)	% Right (S)	Right (C)	% Right (C)
LRC	367	80.84	347	76.43
S-list	333	73.35	318	70.04
BFP	270	59.47	221	48.68
Naive	230	50.66	254	55.95

Figure 1: Pronoun Resolution Algorithms over (S)entences and (C)lauses

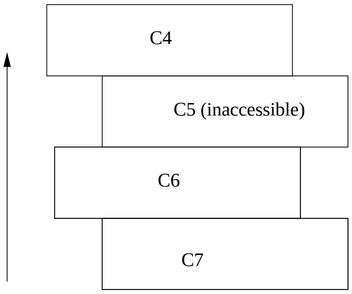

Figure 2: Accessibility due to Clause Embedding

thus many clauses are deemed unfairly inaccessible. To counter this, we developed two transformations to a RST tree: the first involves replacing multi-clausal sentences with one clause in the RST tree; and the second involves merging all subtrees that have a satellite leaf in a relation with a subtree consisting of all leaves, one of which is a nucleus (see Figure 3 (1) for an example).

The intuition with the first transform (SENT) is that many of the errors in the original approximation of G&S based on RST are intrasentential. By merging the clauses together, the tree becomes flattened, and all entities within a sentence are accessible. An example of this transform is in Figure 3 in which one assumes the clauses C1, C2 and C3 of the RST subtree in (1) are constituents of one sentence. Doing the SENT transform yields the result in (3), a subtree that is now a leaf of the sentence reconstructed.

The intuition with the second transform (SAT) is that satellite leaves that modify a nucleus subtree are very closely related to the content of the nucleus leaf of that subtree, as well as the satellite leaf of that subtree. By merging them, the tree is flattened, and pronouns within the original satellite leaf can refer to clauses in the subtree since they are now at the same depth. (2) in Figure 3 provides an illustration of the satellite transformation on (1). The side-effect of this transformation is that the RST tree is no longer binary. Finally, a third transform (SENT-SAT) involves using both of the transforms on the corpus to flatten the tree even more.

In addition, Ide and Cristea note that all exceptions to accessibility in their corpus analysis come from pronouns and antecedents in attribution relations (such as “he said....”). Our corpus exhibits a similar trend: 105 pronouns don’t have an antecedent found in the DRA, out of these 105 inaccessibility cases, 73 are in attribution relations. Though there are 32 unaccounted for (usually because there was an intervening satellite node that prevented reference) the attribution relation tends to be a big block in accessibility still. Another transformation, ATT, is used to counter this by simply merging leaves that stand in attribution relations. So if a subtree has two leaves in an attribution relation, it is replaced by a leaf with the text of the two original leaves merged. This process is similar to SENT.

5 Results

Both algorithms were run over the original RST corpus, ATT (attribution merge) and SAT (satellite merge) transformation of our original corpus (see Figure 4). The (S) version means that the LRC intrasentential search was used over the entire sen-

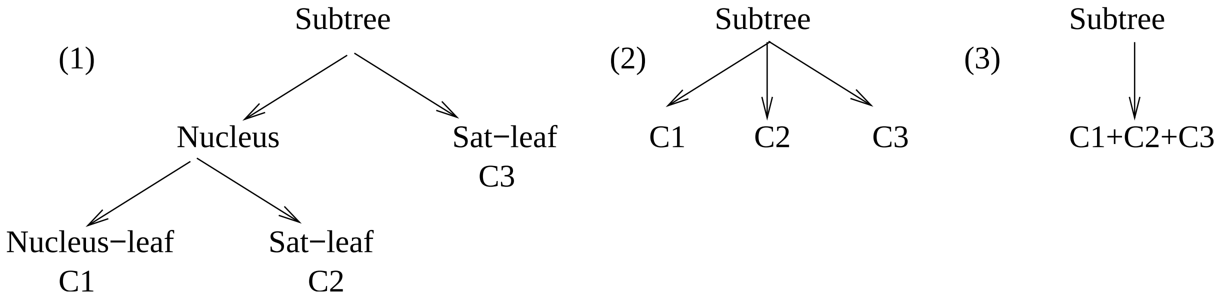

Figure 3: Satellite Transform (2) and Sentence Transform (3)

tence, not just the clause that the pronoun occupies (C). This means that the current sentence is always searched, and if a referent is not found, previous clauses are searched. The (*) signals that the algorithm does not search the previous clause as a default.

Because the SENT transformations created unbalanced RST trees, the Veins algorithm could not be tested with that transform. The results in Figure 5 show how the Grosz and Sidner algorithm fares over the SENT and SENT-SAT transforms with and without using the last-seen metric.

Without the attribution transform, the Veins Algorithm (S) gets only 6 of the 10 pronouns resolved correctly. The G&S algorithms do about as well without segmentation. With the transformations, all the algorithms resolve all 10 cases correctly. However, it should be noted that the original LRC algorithm also resolves all correctly. This success rate is due to the fact that 9 of the 10 pronouns are either “he” or “him” and there are no other candidates with masculine gender in the discourse up to that point. So a simple search through a history-list would resolve these correctly. The other long-distance pronoun is a plural (“their”) and again there are no competing antecedents.

6 Discussion

Discourse decomposition can be evaluated in two ways: intrasentential breakdown (clausal level) and intersentential breakdown (discourse level). In the intrasentential case, all the algorithms performed better when using the (S) method, that is, when the intrasentential search called for searching the sentence the pronoun is in, as opposed to just the clause the pronoun is in. This indicates that ordering clauses by their depth within the sentence or by the Veins information does not improve intrasentential performance, and thus one is better off searching based on grammatical function than incorporating clausal information.

One can evaluate the intersentential decomposition by testing whether the pronouns with long-distance antecedents are resolved correctly. Determining global discourse structure involves finding the middle ground between strict segmentation (using the exact RST tree) and under-segmenting. Too strict a segmentation means that antecedents can be deemed incorrectly inaccessible; very little segmentation means that too many competing antecedents become available since referents are not deemed inaccessible. In our corpus, evaluating intersentential decomposition is difficult because all of the long-distance pronouns have no competing antecedents, so no discourse structure is required to rule out competitors. Therefore it is hard to draw concrete conclusions from the fact G&S on the SENT and SENT-SAT transforms performs the same as LRC algorithm. However, it is promising that this metric does get all of them right, at least it is not overly restrictive. The only way to check if the method under-segments or is a good model is by testing it on a corpus that has long-distance pronouns with competing potential referents. Currently, we are annotating a corpus of dialogs for coreference and rhetorical structure to test this method. It should also be noted that even if an intersentential decomposition method performs the same as knowledge-poor method, it has the advantage of at least decreasing the search space for each pronoun.

Finally, we developed an algorithm for Veins Theory that uses VT to constrain the initial search for

Transform	Veins (S)	Veins (C)	G&S (S*)	G&S (S)	G&S (C)
original	78.85	76.65	72.55	78.90	71.40
ATT	79.30	78.19	73.68	79.30	76.32
SAT	78.85	76.42	73.63	79.08	73.85

Figure 4: Pronoun Resolution Algorithms over ATT and SAT corpora

Transform	G&S(*)	G&S
SENT	78.51	80.84
SENT-SAT	79.74	80.84

Figure 5: Grosz and Sidner over SENT and SENT-SAT corpora

a referent, if one is not found, LRC is used as a default. As suggested by the VT authors, we merged clauses in attribution relations, and this improved performance slightly, but not enough to better 80.84%. VT run on the SAT transform offered no performance enhancement since the theory already makes the nucleus subtrees accessible to satellite leaves.

In conclusion, this study evaluates the theory that clausal segmentation should aid in pronoun resolution by testing two algorithms based on two leading theories of discourse segmentation. Both approaches have the promise of improving pronoun resolution by 1. making search more efficient by blocking utterances or classes from consideration, thus speeding up the search for an antecedent and 2. making search more successful by blocking competing antecedents. We use resolution accuracy for all pronouns and accuracy over long-distance pronominalizations as metrics of success. Our results indicate that incorporating discourse structure does not improve performance, and in most cases can actually hurt performance. However, due to the composition of long-distance pronouns in the corpus, it is necessary to test the G&S algorithm on the SENT transform before drawing a definitive conclusion on the theory.

7 Future Work

Since cases of long distance pronoun resolution are rare, most of the gains in improving accuracy will come from correctly resolving pronouns intrasententially and with the previous utterance. Our error analysis shows that in many cases, determining the coherence relations as Kehler suggests (2002) (such as detecting parallelism between sentences or within sentences) could improve interpretation. In addition, many errors stem from competing antecedents in which incorporating knowledge of the verbs and the entities discussed would prove invaluable.

Finally, our research here has assumed perfect knowledge of discourse structure. Ultimately, the goal is to be able to incrementally build discourse structure while processing a sentence. For this to occur, one has to take into account forms of referring expression, cue words, changes in tense, etc. There has been some work in this area such as Hahn and Strube (1997) who developed an algorithm for building a discourse hierarchy incrementally from changes in theme and centered entities.

Acknowledgements

We would like to thank Mary Swift and three anonymous reviewers for insight and comments. Partial support for the research reported in this paper was provided by the DARPA research grant no. F30602-98-2-0133 to and the ONR grant no. N00014-01-1-1015, both to the University of Rochester.

References

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings, 25th Annual Meeting of the ACL*, pages 155–162.

Donna K. Byron and Joel R. Tetreault. 1999. A flexible architecture for reference resolution. In *Proceedings, 9th Conference of the EACL*, pages 229–232.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Denmark, September.

Dan Cristea, Nancy Ide, and Laurent Romary. 1998. Veins theory: An approach to global cohesion and coherence. In *Proceedings of ACL/COLING*, pages 281–285, Montreal, Canada, August.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. *Proceedings of the Sixth Workshop on Very Large Corpora*.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

Udo Hahn and Michael Strube. 1997. Centered segmentation: Scaling up the centering model to global discourse structure. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 104–111, Somerset, New Jersey.

Nancy Ide and Dan Cristea. 2000. A hierarchical account of referential accessibility. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL 2000*, pages 416–424, Hong Kong, China.

Megumi Kameyama. 1998. Intrasentential centering: A case study. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 89–112. Oxford University Press.

Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8(3):243–281.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Lingusitics*, 19(2):313–330.

R. Mitkov. 2000. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *2nd Discourse Anaphora and Anaphora Resolution Colloquium*, pages 96–107.

M. Moser and J.D. Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

Massimo Poesio and Barbara di Eugenio. 2001. Discourse structure and accessibility. In *ESSLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

Michael Strube. 1998. Never look back: An alternative to centering. In *COLING/ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, volume 2, pages 1251–1257, Montreal, Canada.

Joel R. Tetreault. 1999. Analysis of syntax-based pronoun resolution methods. In *Proceedings, 27th Annual Meeting of the Association for Compuational Linguisitcs*, pages 602–605.

Joel R. Tetreault. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Joel R. Tetreault. 2002. Clausal structure and pronoun resolution. In *4th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 217–220.

Marilyn Walker. 2000. Toward a model of the interaction of centering with global discourse structure. *Verbum*.