Discourse Structure and Co-Reference: An Empirical Study

Dan CRISTEA
Dept. of Computer Science
University "A.I. Cuza"
Iasi, Romania
dcristea@infoiasi.ro

Nancy IDE
Dept. of Computer Science
Vassar College
Poughkeepsie, NY, USA
ide@cs.vassar.edu

Daniel MARCU
Information Sciences Institute
University of Southern California
Los Angeles, CA, USA
marcu@isi.edu

Valentin TABLAN
Dept. of Computer Science
University "A.I. Cuza"
Iasi, Romania
valyt@infoiasi.ro

Abstract

We compare the potential of two classes of linear and hierarchical models of discourse to determine co-reference links and resolve anaphors. The comparison uses a corpus of twenty texts, which were manually annotated for co-reference and discourse structure.

1. Motivation

Most current anaphora resolution systems implement a pipeline architecture with three modules (Lappin and Leass, 1994, Mitkov, 1997, Kameyama 1997):

1. A Collect module determines a list of potential antecedents (LPA) for each anaphor (pronoun, definite noun, proper name, etc.) that have the potential to resolve it.
2. A Filter module eliminates referees incompatible with the anaphor from the LPA.
3. A Preference module determines the most likely antecedent on the basis of an ordering policy.

In most cases, the Collect module determines an LPA by enumerating all antecedents in a window of text that precedes the anaphor under scrutiny (Hobbs, 1978, Lappin and Leass, 1994, Mitkov, 1997, Kameyama, 1997, Ge *et al.,*1998). This window can be as small as two or three sentences or as large as the entire preceding text. The Filter module usually imposes semantic constraints by requiring that the anaphor and potential antecedents have the same number and gender, that selectional restrictions are obeyed, etc. The Preference module imposes preferences on potential antecedents on the basis of their grammatical roles, parallelism, frequency, proximity, etc. In some cases, anaphora resolution systems implement these modules explicitly (Hobbs, 1978, Lappin and Leass, 1994, Mitkov, 1997, Kameyama, 1997). In other cases, these modules are integrated by means of statistical (Ge *et al.,* 1998) or uncertainty reasoning techniques (Mitkov, 1997).

The fact that current anaphora resolution systems rely exclusively on the linear nature of texts in order to determine the LPA of an anaphor seems odd, given that several studies have claimed that there is a strong relation between discourse structure and reference (Sidner, 1981, Grosz and Sidner, 1986, Grosz *et al.*, 1995, Fox, 1987, Vonk *et al.,* 1992, Azzam *et al.*, 1998, Hitzeman and Poesio, 1998). These studies claim, on the one hand, that the use of referents in naturally occurring texts imposes constraints on the interpretation of discourse; and, on the other, that the structure of discourse constrains the LPAs to which anaphors can be

resolved. The oddness of the situation can be explained by the fact that both groups seem *prima facie* to be right. Empirical experiments studies that employ linear techniques for determining the LPAs of anaphors report recall and precision anaphora resolution results in the range of 80% (Lappin and Leass, 1994, Ge *et al.,* 1998). Empirical experiments that investigated the relation between discourse structure and reference also claim that by exploiting the structure of discourse one has the potential of determining correct co-referential links for more than 80% of the referential expressions (Fox, 1987; Cristea *et al*, 1998) although to date, no discourse-based anaphora resolution system has been implemented. Since no direct comparison of these two classes of approaches has been made, it is difficult to determine which group is right, and what method is the best.

In this paper, we attempt to fill this gap by empirically comparing the potential of linear- and hierarchical models of discourse to correctly establish co-referential links in texts, and hence, their potential to correctly resolve anaphors. Since it is likely that both linear- and discourse-based anaphora resolution systems can implement similar Filter and Preference strategies, we focus here only on the strategies that can be used to Collect lists of potential antecedents. Specifically, we focus on determining whether discourse theories can help an anaphora resolution system determine LPAs that are "better" than the LPAs that can be computed from a linear interpretation of texts.

Section 2 outlines the theoretical assumptions of our empirical investigation. Section 3 describes our experiment. We conclude with a discussion of the results.

2. Theoretical background

Our approach is based on the following assumptions:

1. For each anaphor in a text, an anaphora resolution system must produce an LPA that contains a referent to which the anaphor can be resolved. The size of this LPA varies from system to system, depending on the theory a system implements.
2. The smaller the LPA (while retaining a correct antecedent), the less likely that errors in the Filter and Preference modules will affect the ability of a system to select the appropriate referent.
3. Theory A is better than theory B for the task of reference resolution if theory A produces LPAs that contain more antecedents to which anaphors can be correctly resolved than theory B, *and* if the LPAs produced by theory A are smaller than those produced by theory B. For example, if for a given anaphor theory A produces an LPA that contains a referee to which the anaphor can be resolved, while theory B produces an LPA that does not contain such a referee, theory A is better than theory B. Moreover, if for a given anaphor theory A produces an LPA with two referees and theory B produces an LPA with seven referees (each LPA containing a referee to which the anaphor can be resolved), theory A is considered better than theory B because it has a higher probability of solving that anaphor correctly.

We consider two classes of models for determining the LPAs of anaphors in a text:

- ***Linear-k*** *models*

This is a class of linear models in which the LPAs include all the references found in the discourse unit under scrutiny and the ***k*** discourse units that immediately precede it. **Linear-0** models an approach that assumes that all anaphors can be resolved intra-unit; **Linear-1**

models an approach that corresponds roughly to centering (Grosz *et al.,* 1995). **Linear-k** is consistent with the assumptions that underlie most current anaphora resolution systems, which look back ***k*** units in order to resolve an anaphor.

- ***Discourse-VT-k*** *models*

In this class of models, LPAs include all the references found in the discourse unit under scrutiny and the ***k*** discourse units that hierarchically precede it, determined according to *Veins Theory (VT)* (Cristea *et al.,* 1998).

VT extends and formalizes the relation between discourse structure and reference proposed by Fox (1987). It identifies "veins" over discourse structure trees that are built according to the requirements put forth in Rhetorical Structure Theory (RST) (Mann and Thompson, 1987). The fundamental intuition underlying VT is that the RST-specific distinction between nuclei and satellites constrains the range of referents to which anaphors can be resolved; in other words, the nucleus-satellite distinction induces for each anaphor a *domain of referential accessibility* (DRA). In contrast to Centering (Grosz *et al*, 1995), where the DRAs are limited to discourse segments, the DRAs (veins) specific to VT are global: they extend across discourse segment boundaries while restricting the set of units that needs to be considered in order to resolve anaphors (Cristea *et al.* 1998).

For example, according to VT, in order to resolve the anaphor *He* found in unit 3 of the tree in Figure 1, one should look for potential antecedents only in the units that characterize the vein of unit 3: in this case, units 1 and 3. Although unit 2 is linearly closer to unit 3 than 1, VT stipulates that it should not be considered when solving anaphors found in 3, because 2 is a satellite that cannot be accessed from 3. In the tree in Figure 1, nuclei are underlined.

In general, co-referential relations, such as the identity relation that was used in the annotations in our study (see section 3), induce equivalence classes over the set of referential expressions in a text. When hierarchical adjacency is considered, an anaphor may be resolved to a referent that is not the closest in a linear interpretation of a text. However, because referential expressions are organized in equivalence classes, it is sufficient that an anaphor is resolved to *some* member of the set. This is consistent with the distinction between "direct" and "indirect" references discussed by Cristea, *et al.* (1998).

1. *Mr. Casey* 46 years old was president of J&J's Mcneil pharmaceutical subsidiary.
2. The subsidiary was merged with another J&J unit, Ortho Pharmaceutical Corp., this year in a cost-cutting move.
3. *He* succeeds M. James Barrett, 50, as president of genetic therapy.

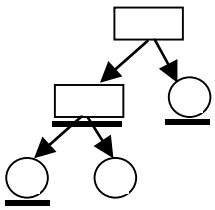

Figure 1 : Short text and RST tree (relations ignored) with veins

3. The Experiment

3.1 Materials

We used twenty newspaper texts that were annotated manually for co-reference relations of identity (Hirschman and Chinchor, 1997). The co-reference relations define equivalence classes on the set of all marked referents in a text. The lengths of the texts varied widely (the mean σ is 408 words and the standard deviation μ is 376). We also used a set of twenty rhetorical structure trees built for the same texts in the style of Mann and Thompson (1988).[1] The average number of elementary discourse units in each text is 52.

[1] The tree construction process is described in Marcu *et al.* (1999).

3.2 Method

The annotations for co-reference relations and rhetorical structure trees for the twenty MUC texts were fused, yielding representations that reflect not only the discourse structure, but also the co-reference equivalence classes specific to each text.[2] The potential of the two classes of models discussed in section 2 (Linear-k and Discourse-VT-k) to correctly establish co-referential links was then evaluated. For each model, each *k,* and each marker referential expression, we determined whether or not the corresponding LPA of size *k* contained a referee from the same equivalence class. For simplicity, and to enable an easier comparison of the models, we approximated each LPA by using the domain of referential accessibility (DRA) that subsumed that LPA. For example, if an LPA for a referential expression in discourse unit 5 contained five antecedents that were realized in units 3, 4, and 5, we considered that the DRA of that expression is given by units 3, 4, and 5. When we compare the models, we rely on the sizes of the DRAs rather than those of the LPAs.

The potential $p(M,a,DRA_k)$ of a model M to determine correct co-referential links with respect to one referential a, given a corresponding DRA of size k (DRA_k) was assigned the value 1 if the DRA contained a co-referent from the same equivalence class as a. Otherwise, $p(M,a,DRA_k)$ was assigned the value 0. The potential $p(M,C,k)$ of a model M to determine correct co-referential links for all referential expressions in a corpus of texts C using DRAs of size k was computed as the sum of the potentials $p(M,a,DRA_k)$ of all referential expressions a in C. This potential was normalized to a value between 0 and 1 by dividing $p(M,C,k)$ by the number of referential expressions in the corpus that had an antecedent.

By examining the potential of each model to correctly determine co-referential expressions for each k, it is possible to determine the degree to which an implementation of a given approach can contribute to the overall efficiency of anaphora resolution systems. That is, if a given model has the potential to determine correctly a large percentage of co-referential expressions with small DRAs, an anaphora resolution system implementing that model will have to consider fewer options overall. Hence, the probability of error is reduced.

4. Results

4.1 Limited VT-k models vs. unlimited Linear-k models

The graph in Figure shows the potentials of the Linear-k and Discourse-VT-k models to correctly determine co-referential links for each k from 1 to 103, which represents the number of elementary discourse units in the largest text. As the graph shows, the potentials initially increase sharply with *k;* Linear models overtake VT models for k=5 (see the enlargement in figure 3). Discourse-VT-k models never reach 100% potential because they are very discriminating: only a small number of units is associated with each DRA. Even for large k, no units are added to DRAs beyond those computed according to VT. As a consequence, it is possible that VT cannot determine correct co-referential links for all referential expressions. In contrast, the Linear model is allowed to search as many units as necessary in order to establish co-referential links (although it is arguable that any system will resolve an anaphor to an antecedent that was introduced, say, 40 units before the unit that contains the anaphor).

[2] The corpus we used contained a total of 30 texts, each annotated for coreference and discourse structure. Because the fusing of the two annotations was done semi-automatically and was time consuming, we did not manage to carry out the experiments described here on the whole corpus. We intend to do so in the final version of this paper.

Figure 4 shows a direct comparison of Linear-k and VT-k potentials. The average value of *k* for which the potentials are equal (VT-k /Linear-k = 1) is 5. The graph in Figure 4 suggests that, on average, looking back a few discourse units using VT rather than a linear model increases the chance that an anaphora resolution system can correctly resolve anaphors. At the maximum value of *k* for which the potentials are equal, there is a potential of resolving 91% of the anaphors.

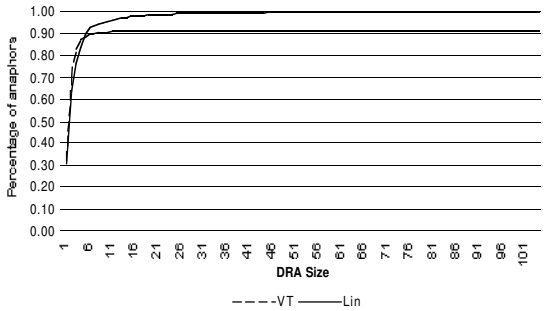

Figure 2 : Potential of Linear-k and Discourse VT-k models to resolve anaphors

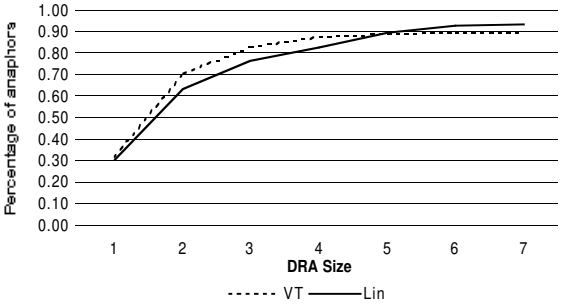

Figure 3 : Potential of Linear-k and Discourse-VT-k models to resolve anaphors, $k < 8$

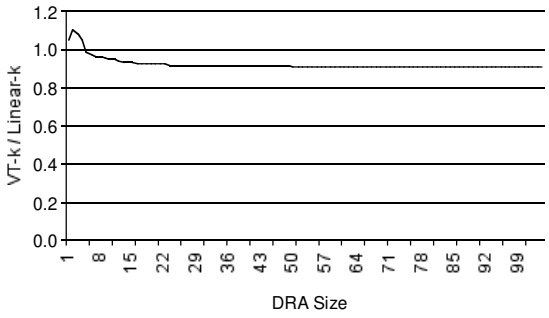

Figure 4 : Direct comparison of Linear-k and Discourse-VT models

We have also investigated how the size *k* – for which discourse and linear theories have the potential of being equally effective in providing lists of possible antecedents to anaphora resolution systems – relates to the size of the texts. The graph in Figure 5, which plots the one-crossing-point for VT-k/Linear-k for each text in the corpus, suggests that the two parameters correlate. More precisely, as the lengths of the texts increase, the size *k* of the DRAs for which the Linear-k models outperform the VT-k models increases as well. For texts that contain as many as 100 units, Linear-k does better than VT-k only for DRAs that contain more than (approximately) 15 units. We believe that most practical anaphora resolution systems rarely look back more than 15 units to choose an antecedent. Therefore, it seems that relying on discourse models to determine lists of potential antecedents of referential expressions can be beneficial.

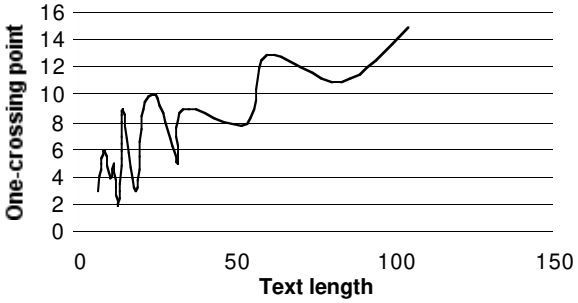

Figure 5 : Correlation of *k* and text length

4.2 Limited VT-k models vs. limited Linear-k models

The results presented so far are somewhat biased because Linear-k models are allowed to extend their DRAs to contain, at the limit, all of the units in a text, while VT-k models are not. Therefore, in order to provide a fairer comparison, we have also examined the case where the size of DRAs for Linear-k models is limited to the maximum number of units in VT-k DRAs (Figures 6, 7, and 8). When this

constraint is imposed, Discourse-VT-k models outperform linear models for all *k* by 5 to 15%.

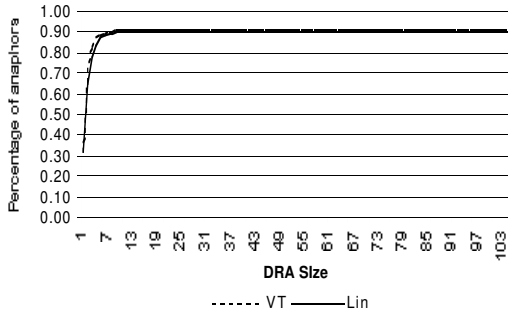

Figure 6 : Linear-k and VT-k potentials (DRA size limited)

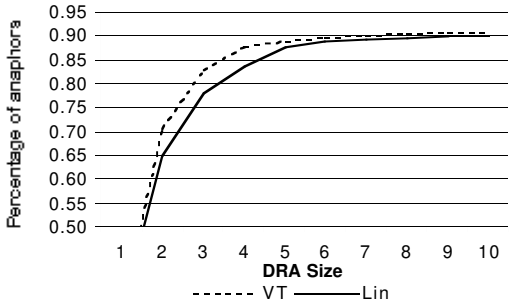

Figure 7 : Linear-k and VT-k potentials (DRA size limited), $k < 8$

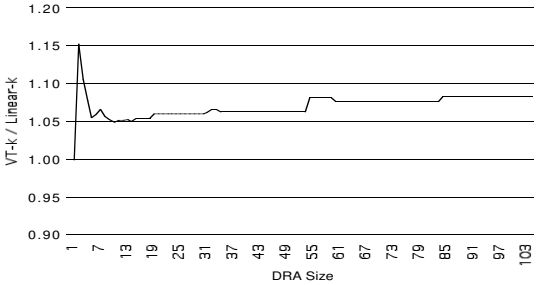

Figure 8 : Direct comparison of Linear-k and Discourse-VT-k (DRA size limited)

4.3 Resolving anaphors

So far, we have investigated Linear-k and Discourse-VT-k models only with respect to their potential to establish correct co-referential links. In order to evaluate their potential to resolve anaphors, we partitioned the set of referents in the texts under scrutiny into four classes, as shown in Table 1.[3]

[3] Since the partitions were determined automatically, it is possible that the anaphor type assignments are not 100% correct. We intend to validate these assignments manually for the final version of the paper.

Expression Type	Occurrences
Definite NPs	162
Pronouns	183
Proper Nouns	262
Indefinite NPs	231
Total	838

Table 1 : Anaphor distribution in the corpus

The referential expressions in Table 1 include only those referents that were co-referential with a referent appearing previously in the text. For each type of referential expression and each *k,* we computed the potential of Linear-k and Discourse-VT-k models to resolve it. Figures 9, 10, 11, and 12 display the direct comparison between the VT-k and Linear-k potentials.

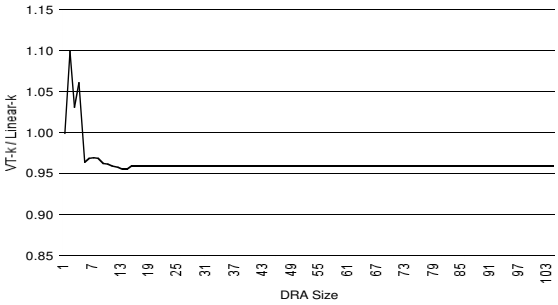

Figure 9 : VT-k / Linear-k for definite NPs

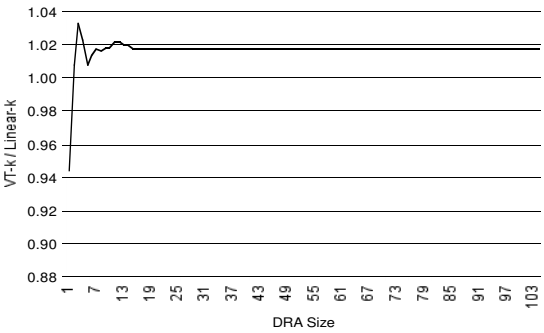

Figure 10 : VT-k / Linear-k for indefinite NPs

On average, Linear-k models outperform VT-k models for *k* greater than 5 and 4 for definite NPs and pronouns, respectively. VT-k models outperform Linear-k models for all *k* for indefinite NPs and proper nouns.

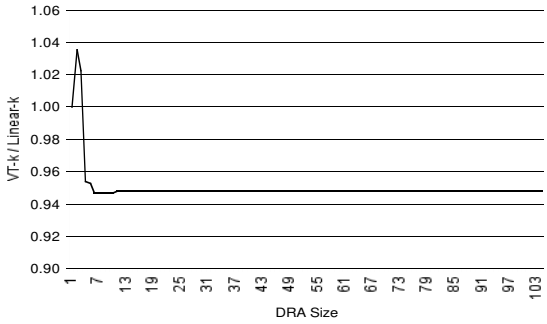

Figure 11 : VT-k / Linear-k for pronouns

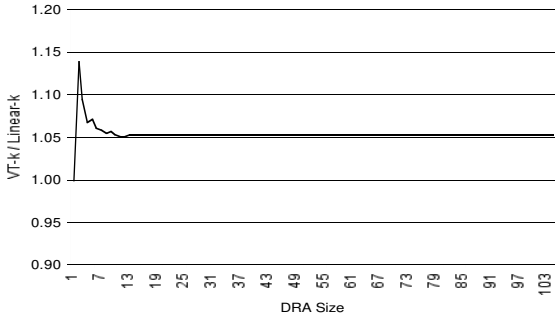

Figure 12 : VT-k / Linear-k for proper nouns

5. Summary

Our results suggest that by exploiting the hierarchical discourse structure of texts one can increase the potential of natural language systems to correctly determine co-referential links and resolve anaphors. In general, the potential to correctly determine co-referential links is greater for VT, the discourse structure model used in our study, than for linear models when one uses DRAs of relatively small size. In addition, it appears that the size of the DRA for which hierarchical discourse models and linear models are equally likely to determine co-referential links directly increases with the size of the text. Hence, the longer the texts, the more likely that hierarchical discourse models can improve the performance of anaphora resolution systems. However, after a certain threshold, the linear models prove to have a better potential for determining co-referential links than hierarchical discourse models. Determining what model to employ and in what circumstances to rely on one model or another is an area that requires further investigation.

Our experiment shows that the potential to determine co-referential links and resolve anaphors increases when the five most recent hierarchically adjacent units, instead of the five most recent linearly adjacent units, are used. Given that 91% of all references in the texts we examined can be resolved using DRAs of size five or less suggests that hierarchical models can contribute significantly to the task of co-reference resolution.

Our current data is for relatively short texts. In the final paper, we will provide data for a larger corpus, including several longer texts, which will enable us to validate our initial findings.

References

Azzam S., K. Humphreys and R. Gaizauskas (1998). Evaluating a Focus-Based Approach to Anaphora Resolution. *Proceedings of COLING-ACL'98,* 74-78.

Cristea, D., N. Ide, and L. Romary (1998). Veins Theory: A Model of Global Discourse Cohesion and Coherence. *Proceedings of COLING-ACL'98,* 281-85.

Fox, B. (1987). *Discourse Structure and Anaphora. Written and Conversational English.* No 48 in *Cambridge Studies in Linguistics*, Cambridge University Press.

Ge N., J. Hale, and E. Charniak (1998). A Statistical Approach to Anaphora Resolution. *Proceedings of the Sixth Workshop on Very Large Corpora,* 161-170.

Grosz, B. and Sidner, C. (1986). Attention, Intention and the Structure of Discourse. *Computational Linguistics,* 12, 175-204.

Grosz, B.J., Joshi, A.K. and Weinstein, S. (1995) Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics,* 12(2), 203-225.

Hahn U. and M. Strübe (1997). Centering in-the-large: Computing referential discourse segments. *Proceedings of ACL-EACL'97*, 104-111.

Hirschman L. and N. Chinchor (1997). MUC-7 Coreference Task Definition.

Hitzeman J., and M. Poesio (1998), Long Distance Pronominalization and Global Focus. *Proceedings of COLING-ACL'98,* 550-556.

Hobbs, J. (1978). Resolving Pronoun References, *Lingua,* 44: 311-338.

Kameyama, M. (1997). Recognizing Referential links: an information extraction perspective. *Proceedings of the ACL/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution,* 46-53.

Lappin S. and H. Leass (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics,* 535-561.

Mann, W.C., Thompson S.A. (1988). Rhetorical structure theory: A theory of text organization, *Text,* 8:3, 243-281.

Marcu D., E. Amorrortu, and M. Romera (1999). Experiments in Constructing a Corpus of Discourse Trees. Submitted for publication.

Mitkov, R. (1997). Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches. *Proceedings of the ACL/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution,* 14-21.

Sidner, C. (1981). Focusing on the Interpretation of Pronouns. *Computational Linguistics,* 7:217-231.

Vonk, W., L. Hustinx, and W. Simons (1992). The Use of Referential Expressions in Structuring Discourse. *Language and Cognitive Processes,* 7 (3-4), 301-333.

Walker, M. (1997). Centering, anaphora resolution, and discourse structure. In Walker, M., A. Joshi, and E. Princes eds., *Centering in Discourse,* Oxford Univrsity Press.