Discourse Structure and Co-Reference: An Empirical Study

Dan Cristea
Department of Computer Science
University "A.I. Cuza"
Iaşi, România
dcristea@infoiasi.ro

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

Daniel Marcu
Information Sciences Institute and
Department of Computer Science
University of Southern California
Los Angeles, CA, USA
marcu@isi.edu

Valentin Tablan
Department of Computer Science
University "A.I. Cuza"
Iaşi, România
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 thirty texts, which were manually annotated for co-reference and discourse structure.

1 Introduction

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 Background

2.1 Assumptions

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 referential expressions found in the discourse unit under scrutiny and the k discourse units that hierarchically precede it. The units that hierarchically precede a given unit are determined according to Veins Theory (VT) (Cristea et al., 1998), which is described briefly below.

2.2 Veins Theory

VT extends and formalizes the relation between discourse structure and reference proposed by Fox (1987). It identifies "veins", i.e., chains of elementary discourse units, over discourse structure trees that are built according to the requirements put forth in Rhetorical Structure Theory (RST) (Mann and Thompson, 1988).

One of the conjectures of VT is that the vein expression of an elementary discourse unit provides a coherent "abstract" of the discourse fragment that contains that unit. As an internally coherent discourse fragment, all anaphors and referential expressions (REs) in a unit must be resolved to referees that occur in the text subsumed by the units in the vein. This conjecture is consistent with Fox's view (1987) that the units that contain referees to which anaphors can be resolved are determined by the nuclearity of the discourse units that precede the anaphors and the overall structure of discourse. According to VT, REs of both satellites and nuclei can access referees of immediately preceding nucleus nodes. REs of nuclei can only access referees of preceding nuclei nodes and of directly subordinated satellite nodes. And the interposition of a nucleus after a satellite blocks the accessibility of the satellite for all nodes that are lower in the corresponding discourse structure (see (Cristea et al., 1998) for a full definition).

Hence, 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 (and each referential expression) a Domain of Referential Accessibility (*DRA*). For each anaphor a in a discourse unit u, VT hypothesizes that a can be resolved by examining referential expressions that were used in a subset of the discourse units that precede u; this subset is called the *DRA* of u. For any elementary unit u in a text, the corresponding *DRA* is computed automatically from the rhetorical representation of that text in two steps:

1. **Heads** for each node are computed bottom-up over the rhetorical representation tree. Heads of elementary discourse units are the units themselves. Heads of internal nodes, i.e., discourse spans, are computed by taking the union of the heads of the immediate child nodes that are nuclei. For example, for the text in Figure 1, whose rhetorical structure is shown in Figure 2, the head of span [5,7] is unit 5 because the head of the immediate nucleus, the elementary unit 5, is 5. However, the head of span [6,7] is the list $\langle 6,7 \rangle$ because both immediate children are nuclei of a multinuclear relation.

2. Using the results of step 1, **Vein** expressions are computed top-down for each node in the tree. The vein of the root is its head. Veins of child nodes are computed recursively according to the rules described by Cristea et al.(1998). The *DRA* of a unit u is given by the units in the vein that precede u.

   For example, for the text and RST tree in Figures 1 and 2, the vein expression of unit 3, which contains units 1 and 3, suggests that anaphors from unit 3 should be resolved only to referential expressions in units 1 and 3. Because unit 2 is a satellite to unit 1, it is considered to be "blocked" to referential links from unit 3. In contrast, the *DRA* of unit 9, consisting of units 1, 8, and 9, reflects the intuition that anaphors from unit 9 can be resolved only to referential expressions from unit 1, which is the most important unit in span [1,7], and to unit 8, a satellite that immediately precedes unit 9. Figure 2 shows the heads and veins of all internal nodes in the rhetorical representation.

```
1.  Michael D. Casey, a top Johnson&Johnson
    manager, moved to Genetic Therapy Inc.,
    a small biotechnology concern here,
2.  to become its president and chief
    operating officer.
3.  Mr. Casey, 46 years old, was president of
    J&J's McNeil Pharmaceutical subsidiary,
4.  which was merged with another J&J unit,
    Ortho Pharmaceutical Corp., this year in
    a cost-cutting move.
5.  Mr. Casey succeeds M. James Barrett, 50,
    as president of Genetic Therapy.
6.  Mr. Barrett remains chief executive officer
7.  and becomes chairman.
8.  Mr. Casey said
9.  he made the move to the smaller company
10. because he saw health care moving toward
    technologies like the company's gene therapy
    products.
11. I believe that the field is emerging and is
    prepared to break loose,
12. he said.
```

Figure 1: An example of text and its elementary units. The referential expressions surrounded by boxes and ellipses correspond to two distinct co-referential equivalence classes. Referential expressions surrounded by boxes refer to *Mr. Casey*; those surrounded by ellipses refer to *Genetic Therapy Inc.*.

2.3 Comparing models

The premise underlying our experiment is that there are potentially significant differences in the size of the search space required to resolve referential expressions when using Linear models vs. Discourse-VT models. For example, for text and the RST tree in Figures 1 and 2, the Discourse-VT model narrows the search space required to resolve the anaphor *the smaller company* in unit 9. According to VT, we look for potential antecedents for *the smaller company* in the *DRA* of unit 9, which lists

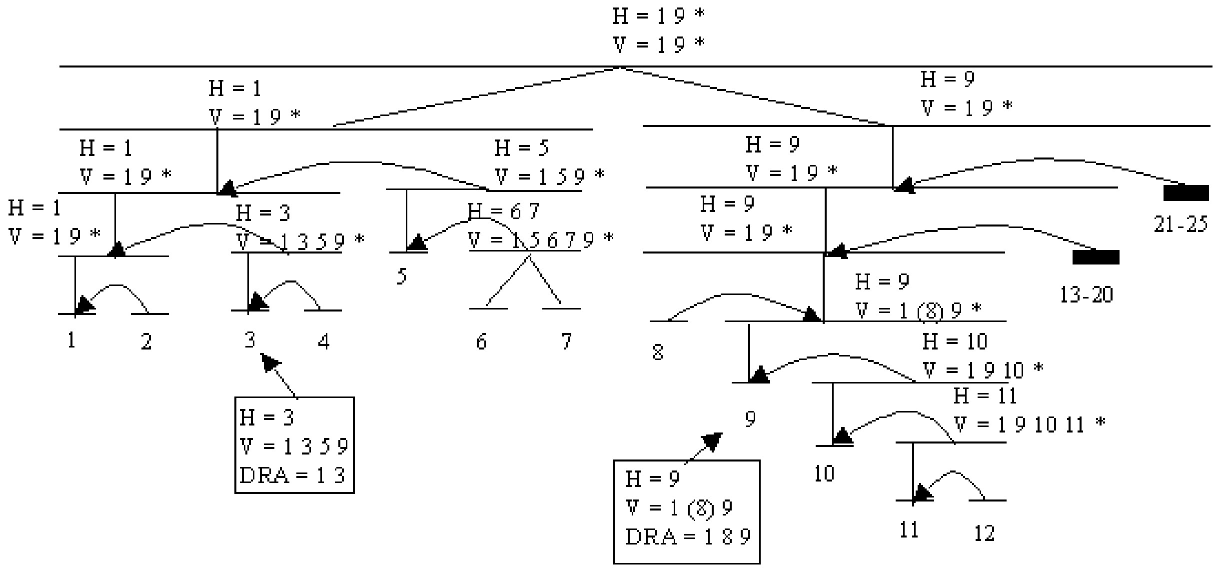

Figure 2: The RST analysis of the text in figure 1. The tree is represented using the conventions proposed by Mann and Thompson (1988).

units 1, 8, and 9. The antecedent *Genetic Therapy, Inc.* appears in unit 1; therefore, using VT we search back 2 units (units 8 and 1) to find a correct antecedent. In contrast, to resolve the same reference using a linear model, four units (units 8, 7, 6, and 5) must be examined before *Genetic Therapy* is found. Assuming that referential links are established as the text is processed, *Genetic Therapy* would be linked back to pronoun *its* in unit 2, which would in turn be linked to the first occurrence of the antecedent,*Genetic Therapy, Inc.,* in unit 1, the antecedent determined directly by using VT.

In general, 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. Similarly, a referential expression can be linked to a referee that is not the closest in a linear interpretation of a text. However, this does not create problems because we are focusing here only on co-referential relations of identity (see section 3). Since these relations induce equivalence classes over the set of referential expressions in a text, it is sufficient that an anaphor or referential expression is resolved to any of the members of the relevant equivalence class. For example, according to VT, the referential expression *Mr. Casey* in unit 5 in Figure 1 can be linked directly only to the referee *Mr Casey* in unit 1, because the *DRA* of unit 5 is $\{1,5\}$. By considering the co-referential links of the REs in the other units, the full equivalence class can be determined. This is consistent with the distinction between "direct" and "indirect" references discussed by Cristea, et al.(1998).

3 The Experiment

3.1 Materials

We used thirty newspaper texts whose lengths varied widely; the mean σ is 408 words and the standard deviation μ is 376. The texts 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 texts were also manually annotated with discourse structures built in the style of Mann and Thompson (1988). Each analysis yielded an average of 52 elementary discourse units. Details of the discourse annotation process are given in (Marcu et al., 1999).

3.2 Comparing potential to establish co-referential links

3.2.1 Method

The annotations for co-reference relations and rhetorical structure trees for the thirty texts were fused, yielding representations that reflect not only the discourse structure, but also the co-reference equivalence classes specific to each text. Based on this information, we evaluated the potential of each of the two classes of models discussed in section 2 (Linear-k and Discourse-VT-k) to correctly establish co-referential links as follows: For each model, each k, and each marked referential expression a, we determined whether or not the corresponding LPA (defined over k elementary units) contained a referee from the same equivalence class. For example, for the Linear-2 model and referential expression *the smaller company* in unit 9, we estimated whether a co-referential link could be established between *the smaller company* and another referential expression in units 7, 8, or 9. For the Discourse-VT-2 model and the same referential expression, we estimated whether a co-referential link could be established between *the smaller company* and another referential expression in units 1, 8, or 9, which correspond to the *DRA* of unit 9.

To enable a fair comparison of the two models, when k is larger than the size of the *DRA* of a given unit, we extend that *DRA* using the closest units that precede the unit under scrutiny and are not already in the *DRA*. Hence, for the Linear-3 model and the referential expression *the smaller company* in unit 9, we estimate whether a co-referential link can be established between *the smaller company* and another referential expression in units 6, 7, 8, or 9. For the Discourse-VT-3 model and the same referential expression, we estimate whether a co-referential link can be established between *the smaller company* and another referential expression in units 1, 8, 9, or 7, which correspond to the *DRA* of unit 9 (units 1, 8, and 9) and to unit 7, the closest unit preceding unit 9 that is not in its *DRA*.

For the Discourse-VT-k models, we assume that the Extended *DRA* (*EDRA*) of size k of a unit u ($\mathit{EDRA}_k(u)$) is given by the first $l \leq k$ units of a sequence that lists, in reverse order, the units of the *DRA* of u plus the $k - l$ units that precede u but are not in its *DRA*. For example, for the text in Figure 1, the following relations hold: $\mathit{EDRA}_0(9) = 9$; $\mathit{EDRA}_1(9) = 9, 8$; $\mathit{EDRA}_2(9) = 9, 8, 1$; $\mathit{EDRA}_3(9) = 9, 8, 1, 7$; $\mathit{EDRA}_4(9) = 9, 8, 1, 7, 6$. For Linear-k models, the $\mathit{EDRA}_k(u)$ is given by u and the k units that immediately precede u.

The potential $p(M, a, \mathit{EDRA}_k)$ of a model M to determine correct co-referential links with respect to a referential expression a in unit u, given a corresponding *EDRA* of size k ($\mathit{EDRA}_k(u)$), is assigned the value 1 if the *EDRA* contains a co-referent from the same equivalence class as a. Otherwise, $p(M, a, \mathit{EDRA}_k)$ is 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 *EDRA*s of size k, is computed as the sum of the potentials $p(M, a, \mathit{EDRA}_k)$ of all referential expressions a in C. This potential is normalized to a value between 0 and 1 by dividing $p(M, C, k)$ by the number of referential expressions in the corpus that have 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 correctly determine a significant percentage of co-referential expressions with small *DRA*s, an anaphora resolution system implementing that model will have to consider fewer options overall. Hence, the probability of error is reduced.

3.2.2 Results

The graph in Figure 3 shows the potentials of the Linear-k and Discourse-VT-k models to correctly determine co-referential links for each k from 1 to 20. The graph in Figure 4 represents the same potentials but focuses only on ks in the interval [2,9]. As these two graphs show, the potentials increase monotonically with k, the VT-k models always doing better than the Linear-k models. Eventually, for large ks, the potential performance of the two models converges to 100%.

The graphs in Figures 3 and 4 also suggest resolution strategies for implemented systems. For example, the graphs suggests that by choosing to work with *EDRA*s of size 7, a discourse-based system has the potential of resolving more than 90% of the co-referential links in a text correctly. To achieve the same potential, a linear-based system needs to look back 8 units. If a system does not look back at all and attempts to resolve co-referential links only within the unit under scrutiny ($k = 0$), it has the potential to correctly resolve about 40% of the co-referential links.

To provide a clearer idea of how the two models differ, Figure 5 shows, for each k, the value of the Discourse-VT-k potentials divided by the value of the Linear-k potentials. For $k = 0$, the potentials of both models are equal because both use only the unit in focus in order to determine co-referential links. For $k = 1$, the Discourse-VT-1 model is about 7%

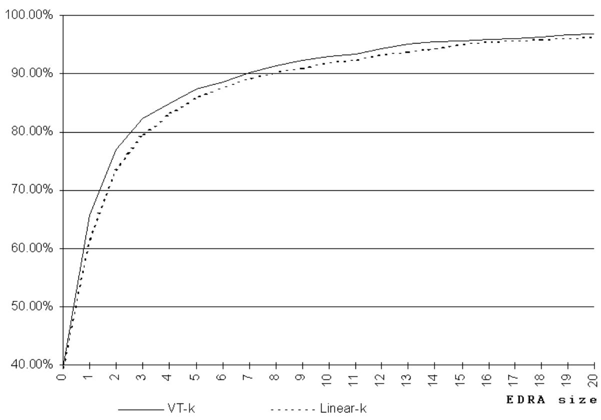

Figure 3: The potential of Linear-k and Discourse-VT-k models to determine correct co-referential links $(0 \leq k \leq 20)$.

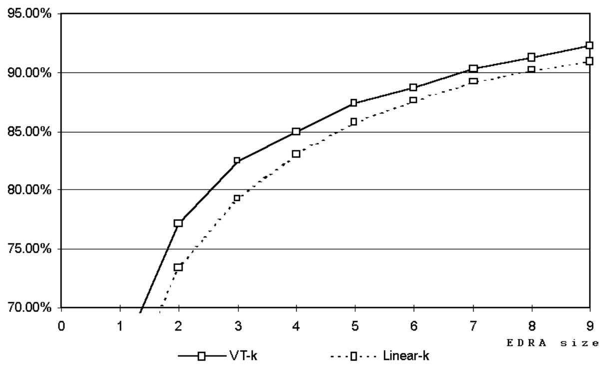

Figure 4: The potential of Linear-k and Discourse-VT-k models to determine correct co-referential links $(2 \leq k \leq 9)$.

better than the Linear-1 model. As the value of k increases, the value Discourse-VT-k/Linear-k converges to 1.

In Figures 6 and 7, we display the number of exceptions, i.e., co-referential links that Discourse-VT-k and Linear-k models cannot determine correctly. As one can see, over the whole corpus, for each $k \leq 3$, the Discourse-VT-k models have the potential to determine correctly about 100 more co-referential links than the Linear-k models. As k increases, the performance of the two models converges.

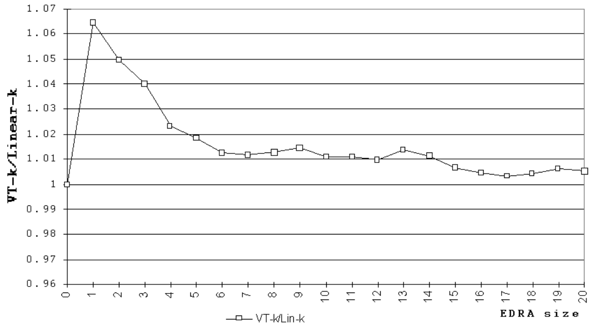

Figure 5: A direct comparison of Discourse-VT-k and Linear-VT-k potentials to correctly determine co-referential links $(0 \leq k \leq 20)$.

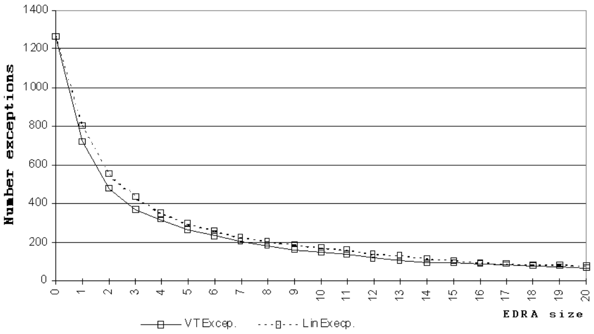

Figure 6: The number of co-referential links that cannot be correctly determined by Discourse-VT-k and Linear-k models $(0 \leq k \leq 20)$.

3.2.3 Statistical Significance

In order to assess the statistical significance of the difference between the potentials of the two models to establish correct co-referential links, we carried out a Paired-Samples T Test for each k. In general, a Paired-Samples T Test checks whether the mean of casewise differences between two variables differs from 0. For each text in the corpus and each k, we determined the potentials of both VT-k and Linear-k models to establish correct co-referential links in that text. For ks smaller than 4, the difference in potentials was statistically significant. For example, for $k = 3, t = 3.345, df = 29, P = 0.002$. For values of k larger than or equal to 4, the difference was no longer significant. These results are consistent with the graphs shown in Figure 3 to 7, which all show that the potentials of Discourse-VT-k and Linear-k models converges to the same value as the value of k increases.

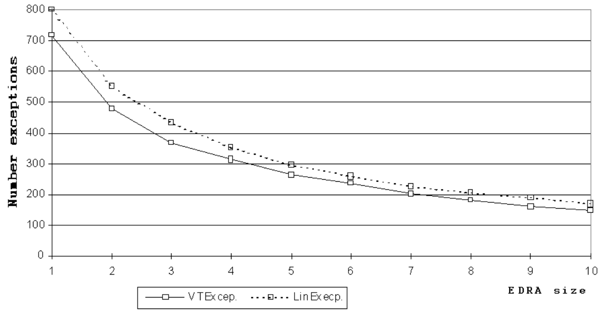

Figure 7: The number of co-referential links that cannot be correctly determined by Discourse-VT-k and Linear-k models $(1 \leq k \leq 10)$.

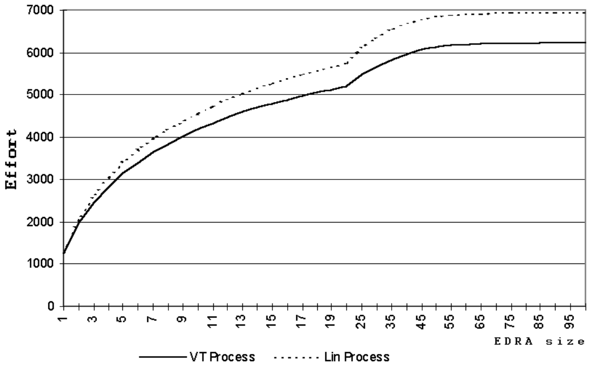

Figure 8: The effort required by Linear-k and Discourse-VT-k models to determine correct co-referential links $(0 \leq k \leq 100)$.

3.3 Comparing the effort required to establish co-referential links

3.3.1 Method

The method described in section 3.2.1 estimates the potential of Linear-k and Discourse-VT-k models to determine correct co-referential links by treating *EDRA*s as sets. However, from a computational perspective (and presumably, from a psycholinguistic perspective as well) it also makes sense to compare the *effort* required by the two classes of models to establish correct co-referential links. We estimate this effort using a very simple metric that assumes that the closer an antecedent is to a corresponding referential expression in the *EDRA*, the better. Hence, in estimating the effort to establish a co-referential link, we treat *EDRA*s as ordered lists. For example, using the Linear-9 model, to determine the correct antecedent of the referential expression *the smaller company* in unit 9 of Figure 1, it is necessary to search back through 4 units (to unit 5, which contains the referent *Genetic Therapy*). Had unit 5 been *Mr. Cassey succeeds M. James Barrett, 50,* we would have had to go back 8 units (to unit 1) in order to correctly resolve the RE *the smaller company*. In contrast, in the Discourse-VT-9 model, we go back only 2 units because unit 1 is two units away from unit 9 ($\mathit{EDRA}_9(9) = 9, 8, 1, 7, 6, 5, 4, 3, 2$).

We consider that the effort $e(M, a, \mathit{EDRA}_k)$ of a model M to determine correct co-referential links with respect to one referential a in unit u, given a corresponding *EDRA* of size k ($\mathit{EDRA}_k(u)$) is given by the number of units between u and the first unit in $\mathit{EDRA}_k(u)$ that contains a co-referential expression of a.

The effort $e(M, C, k)$ of a model M to determine correct co-referential links for all referential expressions in a corpus of texts C using *EDRA*s of size k was computed as the sum of the efforts $e(M, a, \mathit{EDRA}_k)$ of all referential expressions a in C.

3.3.2 Results

Figure 8 shows the Discourse-VT-k and Linear-k efforts computed over all referential expressions in the corpus and all ks. It is possible, for a given referent a and a given k, that no co-referential link exists in the units of the corresponding EDRA_k. In this case, we consider that the effort is equal to k. As a consequence, for small ks the effort required to establish co-referential links is similar for both theories, because both can establish only a limited number of links. However, as k increases, the effort computed over the entire corpus diverges dramatically: using the Discourse-VT model, the search space for co-referential links is reduced by about 800 units for a corpus containing roughly 1200 referential expressions.

3.3.3 Statistical significance

A Paired-Samples T Test was performed for each k. For each text in the corpus and each k, we determined the effort of both VT-k and Linear-k models to establish correct co-referential links in that text. For all ks the difference in effort was statistically significant. For example, for $k = 7$, we obtained the values $t = 3.51, df = 29, P = 0.001$. These results are intuitive: because *EDRA*s are treated as ordered lists and not as sets, the effect of the discourse structure on establishing correct co-referential links is not diminished as k increases.

4 Conclusion

We analyzed empirically the potentials of discourse and linear models of text to determine co-referential links. Our analysis suggests that by exploiting the hierarchical structure of texts, one can increase the potential of natural language systems to correctly determine co-referential links, which is a requirement for correctly resolving anaphors. If one treats all discourse units in the preceding discourse equally, the increase is statistically significant only when a discourse-based coreference system looks back at most four discourse units in order to establish co-referential links. However, if one assumes that proximity plays an important role in establishing co-referential links and that referential expressions are more likely to be linked to referees that were used recently in discourse, the increase is statistically significant no matter how many units a discourse-based co-reference system looks back in order to establish co-referential links.

Acknowledgements. We are grateful to Lynette Hirschman and Nancy Chinchor for making available their corpus of co-reference annotations. We are also grateful to Graeme Hirst for comments and feedback on a previous draft of this paper.

References

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1998. Evaluating a focus-based approach to anaphora resolution. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (COLING/ACL'98)*, pages 74–78, Montreal, Canada, August 10–14.

Dan Cristea, Nancy Ide, and Laurent Romary. 1998. Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (COLING/ACL'98)*, pages 281–285, Montreal, Canada, August.

Barbara Fox. 1987. *Discourse Structure and Anaphora*. Cambridge Studies in Linguistics; 48. Cambridge University Press.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, Montreal, Canada, August 15-16.

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

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, June.

Lynette Hirschman and Nancy Chinchor, 1997. *MUC-7 Coreference Task Definition*, July 13.

Janet Hitzeman and Massimo Poesio. 1998. Long distance pronominalization and global focus. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (COLING/ACL'98)*, pages 550–556, Montreal, Canada, August.

Jerry H. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Megumi Kameyama. 1997. Recognizing referential links: An information extraction perspective. In *Proceedings of the ACL/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, pages 46–53.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

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

Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL'99 Workshop on Standards and Tools for Discourse Tagging*, University of Maryland, June 22.

Ruslan Mitkov. 1997. Factors in anaphora resolution: They are not the only things that matter. a case study based on two different approaches. In *Proceedings of the ACL/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, pages 14–21.

Candace L. Sidner. 1981. Focusing for interpretation of pronouns. *Computational Linguistics*, 7(4):217–231, October–December.

Wietske Vonk, Lettica G.M.M. Hustinx, and Wim H.G. Simons. 1992. The use of referential expressions in structuring discourse. *Language and Cognitive Processes*, 7(3,4):301–333.