A Corpus Study of Referential Choice: The Role of Rhetorical Structure

Andrej A. Kibrik (Institute of Linguistics, Russian Academy of Sciences)
Olga N. Krasavina (Moscow State University and Humboldt University of Berlin)
kibrik@comtv.ru; krasavio@rz.hu-berlin.de

Abstract

This study shares the view that reference in discourse is influenced by the distance to prior mentions of the referent in the discourse. Kibrik (1996, 1999) suggested a measurement of rhetorical distance to assess this factor. In this paper we address three complications created by that methodology when applied to a large corpus of written newspaper texts. These problems include: difference between symmetrical and asymmetrical discourse structures as sites of antecedents; type of rhetorical relations as a factor or referential choice; and multiple (competing) antecedents. In this study we outline a model that is relevant for both theory of referential choice in discourse and applied explorations of anaphora resolution or generation.

1. Introduction: "Will rhetorical structure redeem us?"

Pronominal anaphora has been one of the most favorite study subjects of diverse theoretical frameworks over the years. A growing number of studies in anaphora, especially anaphora resolution, are a characteristic feature of the last decade. However, at the present moment both theoretical and applied approaches are facing a sort of stagnation. Anaphora theorists still do not have a model capable of explaining and/or predicting the use of basic anaphoric devices, validated on large natural language corpora, and computational linguists do not get any significant improvements in their resolution algorithms.

Hierarchical, or rhetorical, structure of discourse is a possibly important but still not sufficiently studied factor that impacts the use of anaphoric devices. There is some evidence for this (Fox 1987, Grosz and Sidner 1986, etc.), the quote in the title is but one of the cries from the heart in the anaphora research community (see Wolters 2001). Unfortunately, the existing heuristics of anaphoric devices use are often too rough (cf. Cristea et al. 1998, 2000). The present study attempts to solve this problem by investigating the following discourse structural features:

- distance to the antecedent
- semantic types of rhetorical relation between clauses
- choice between two or more potential antecedents.

In this study we approach referential phenomena from the production perspective. That is, we are interested in referential choice by the speaker rather than reference resolution. This study builds on a corpus of newspaper American English – the RST Discourse Treebank (Carlson et al. 2003) annotated for rhetorical structure (following the principles of Rhetorical Structure Theory, see Mann and Thompson 1988).

In Section 2, a terse description of Rhetorical Structure Theory follows. Section 3 discusses the role of rhetorical structure in referential choice. In order to evaluate the role of discourse structure we employ the parameter of rhetorical distance proposed in Kibrik (1996, 1999) (section 4). A number of improvements to the prior procedure are discussed in sections 5 to 7. Section 8 concludes this communication.

2. Rhetorical Structure Theory

Rhetorical Structure Theory, or RST (Mann and Thompson 1988), is one of the most widely-used tools to assess discourse coherence, thus bringing a global and local structure of discourse together. According to the RST, discourse is divided into discourse units. Elementary discourse units essentially coincide with clauses. The RST assumes a number of rhetorical relations between discourse units which can be either symmetrical (multinuclear) or asymmetrical (mononuclear). An asymmetrical relation connects a nucleus and a satellite, and a symmetrical relation connects two or more nuclei. Rhetorical relations resemble semantic relations between the main and adjunct clauses in complex sentences, but extend to the discourse level, that is, can connect discourse units irrespective

of sentence boundaries. They are construed as being motivated by the speaker's communicative goals, rather than by the principles of syntax; hence the term "rhetorical".

Elementary discourse units are connected by rhetorical relations into higher order units, among which the same kind of rhetorical relations hold. Discourse units hierarchically grow all the way to the level of discourse global structure, and eventually the discourse as a whole. Discourse units and relations between them are represented as a graph. Each discourse unit is a node of the graph, and rhetorical relations mark the ribs of the graph. Examples of rhetorical graphs will be provided below.

RST explicitly recognizes the inherent ambiguity of discourse, and possibility of various interpretations. Thus rhetorical graphs constructed for a certain text can vary in details. However, if trained analysts are employed, the degree of such variation is very limited. As we use the corpus that was consistently analyzed by a trained team we do not address this problem below.

In the rest of this paper, we are going to discuss what is the impact of rhetorical structure referential choice.

3. Discourse structure and referential choice

From the cognitive-functional point of view, pronominalization correlates with a higher degree of referent's activation in the speaker's cognitive system. In Givón (1983), such activation, or 'accessibility', was assessed by measuring referential distance, i.e. distance in clauses between the anaphor- and antecedent-containing clauses. The distance of 1 was considered to be a sign of high accessibility of a referent, and the most accessible referents were the most probable candidates for pronominalization.

In a number of later works, it has been demonstrated that a distance-based approach can be reinforced if one uses not just the linear distance but distance computed with respect to rhetorical structure (Fox 1987, Kibrik 1996 i.a.). Kibrik used the term 'rhetorical distance' for such method of distance computation. Rhetorical distance thus measures the closeness/remoteness of an anaphor to/from its antecedent in the hierarchical structure of discourse.

There are various ways to compute rhetorical distance. For example, one can compute distance in symmetrical and asymmetrical structures identically or differently. Besides this, there are purely technical or individual matters that can affect the measurement heuristics. Therefore, the method of calculating rhetorical distance should be adjusted to the existing rhetorical annotation in the corpus before it can be applied properly.

It must be kept in mind that genre peculiarities may affect referential choice (Toole 1996). Generally, short sentences, clear argumentation lines, simplified structure are characteristic of newspaper texts. However, some differentiation may be due to variation between "sub-genres": for example, financial reports are different from commentaries on some events – the former are more concise and poor in human referents, the latter present a wider diversity of rhetorical relations and contain a lot of human referents, as well as abstract ones. All examples considered below are commentaries, for the sake of consistency.

4. The approach of Kibrik (1996, 1999) to referential choice and its application to the RST corpus

Kibrik (1996, 1999) proposed a multi-factorial approach to referential choice: up to a dozen factors were identified that influence referential choice, and their relative weights were determined. Rhetorical distance (RhD) turned out the strongest factor of all, even though no single factor, for sure, can explain such a complex phenomenon as referential choice.

The basic method to compute RhD was as follows:

- move along the graph towards the nearest antecedent and count how many horizontal jumps you make

In many cases RhD does not differ from plain linear distance (LinD), see for example graphs in Fig. 1 in which both RhD and LinD from C to A are 2.

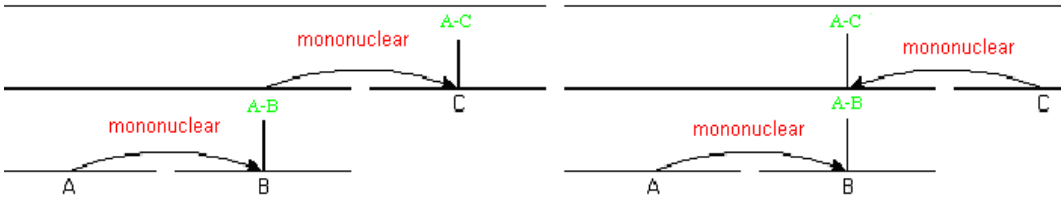

Fig. 1

Now, there are many instances in which RhD does differ from LinD. For example, linear and rhetorical distances from C to A and B in Fig. 2 are different, as shown in Table 1.

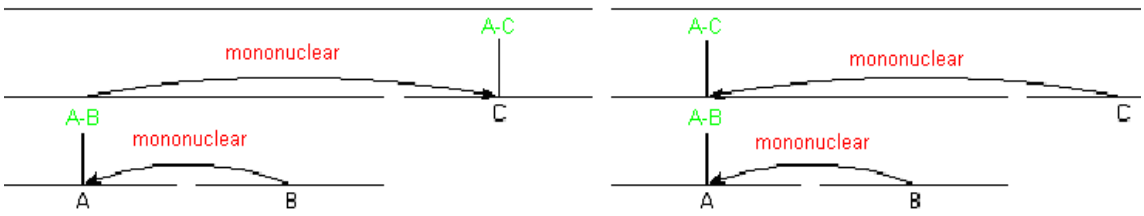

Fig. 2

	From C to B	From C to A
LinD	1	2
RhD	2	1

Table 1.

Rhetorical distance thus captures the hierarchical closeness of nodes that are immediately connected in the discourse structure but separated linearly, as C and A in both configurations in Fig. 2. It also captures the hierarchical separation of linearly adjacent nodes, such as C and B in Fig. 2.

To take a real example from the RST corpus, in Figure 3 node 127 contains two full NPs, Mr. Schaeffer and the couple that serve as antecedents of further referential expressions. Once we encounter anaphoric pronouns in 128, 129, and 130, we search along the graph for the nearest antecedent clause that invariably is 127. Rhetorical distances, according to Kibrik (1996, 1999), are as follows:

- for them in 128 RhD=1
- for he in 129 RhD=1
- for them in 130 RhD=2

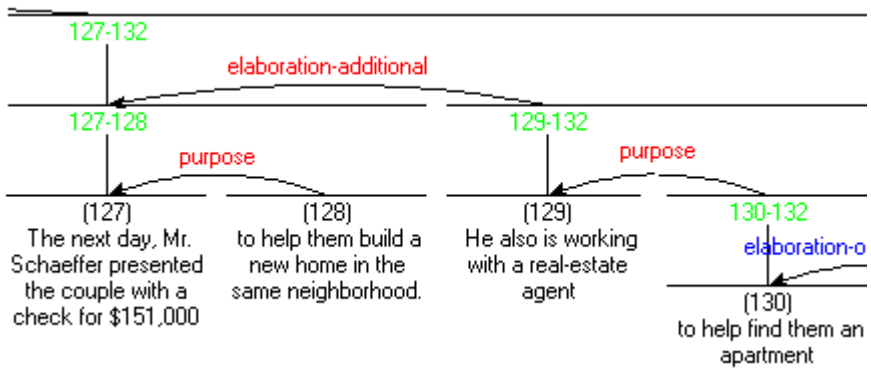

Fig. 3.

A number of other conventions were used in Kibrik (1996, 1999) for computing RhD. When we applied that methodology to the RST corpus we encountered several persisting problems. These problems were acknowledged already in Kibrik's prior work but were not significant there because of the small size of the data set and the genre properties of the text: it was a narrative which is rhetorically quite monotonous anyway (see Kibrik 2002), and it was also very homogenic as a story by one author.

The major problems we encountered in the RST corpus study are the following.

1) It seems necessary to adopt some conventions on how to treat symmetrical structures in the computation of RhD. Empirical facts suggest that antecedents inside symmetrical groups in certain cases are somewhat less accessible compared to antecedents in the nuclei of asymmetrical groups.

2) In some cases plain rhetorical distance seems too coarse a measurement. Certain rhetorical relations make the nodes they connect in a way closer than those connected by other relations. In other words, some relations provide a tighter connection than other.

3) Sometimes there is more than one candidate for the rhetorical antecedent of an anaphor, and a formal procedure to choose one is in order.

Below we consider these three problems one at a time.

5. Treatment of symmetrical relations

In the structures shown above all rhetorical relations are asymmetrical, and this is when the contribution of the rhetorical distance is particularly clear. However, when one deals with symmetrical structures it is not immediately obvious how to compute RhD. Consider the configurations in Fig. 4.

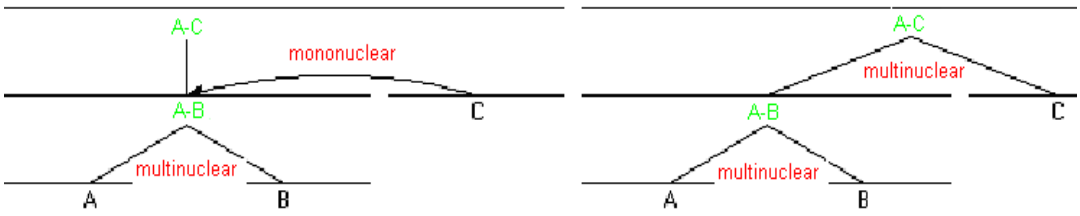

Fig. 4.

What is the distance from C to A and B? Is it the same or different? And is it the same as in the structures in Fig. 1 or not? (Note that one slanting line in symmetrical constructions counts as a vertical jump and thus does not affect the RhD. Only going along a pair of slanting lines, such as from B to A, constitutes a horizontal jump.)

In Kibrik (1996, 1999) a convention was adopted, according to which the RhD from C to B in Fig. 4 is 1 and from C to A is 2. That is, a node connected to a symmetrical group was reinterpreted as the satellite of its linearly last member, and both structures in Fig. 3 were reinterpreted as shown in Fig. 5.

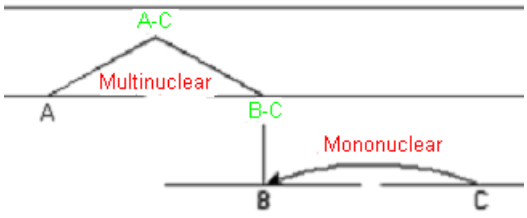

Fig. 5.

However, there is evidence that structures in Fig. 4 and Fig. 5 are not equivalent in terms of antecedent accessibility. This is easy to demonstrate with a couple of made-up examples. Example (1) illustrates the configuration in Fig. 5:

(1) John was playing and Bill was watching a movie because he was through with his homework.

Here the pronoun he has the clear antecedent Bill in clause B despite the theoretically possible interference of another antecedent in clause A. This is due to the different rhetorical distance from C to A and from C to B. Now, example (2) illustrates the configuration in Fig. 4 (either one of the two represented in Fig. 4):

(2) John was playing and Bill was reading. Then he stood up and walked out.

Here it is not quite clear who the pronoun he refers to, and, therefore, a speaker is unlikely to use a structure as in (2). Apparently, both antecedents are in principle accessible due to the symmetrical character of the rhetorical structure, therefore the ambiguity becomes very real and influences the felicity of pronominalization. In (2), the antecedent Bill perhaps has some privilege over John, but this is due to its linear proximity to the anaphor; note that the linear distance remains as a relevant factor in our system, despite its secondary status compared to the rhetorical distance. So, the convention of Kibrik (1996, 1999) actually smuggled linear proximity into the RhD measurement, and as a result linear distance was taken into account twice.

Thus the prior convention must be replaced. It is not a good idea to simply postulate that both A and B in Fig. 3 are as accessible as are the nuclei B in Fig. 1 and A in Fig. 2. We hypothesize that the single nucleus of an asymmetrical structure (Fig. 1, 2) is more accessible as an antecedent clause than the multiple nuclei of symmetrical structures (Fig. 4). We propose the following convention:

- when we penetrate into a symmetrical structure, a penalty of 0.5 is added

Thus, for example, the rhetorical distance from C to either A or B in Fig. 4 will be 1.5. In the implementation of our model of referential choice we are going to explore this hypothesis and see if it improves the performance of the model.

Likewise we propose to charge a 0.5 penalty when the anaphor itself is in a symmetrical structure, for getting out of such structure. A number of further issues remain to be resolved. In particular, there may be more than one hierarchical layer of symmetrical structures, see Fig. 6.

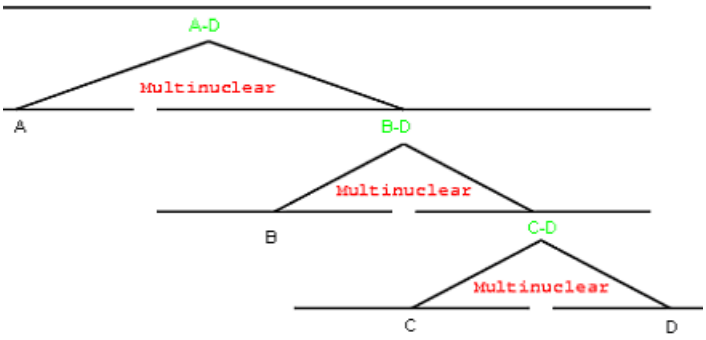

Fig. 6.

The rhetorical distance from C to B is 1.5, that is, 0.5 for getting out of the symmetrical structure C-D plus 1 for one horizontal jump (from C-D to B). The question is what is the RhD from C to A: also 1.5 or 2? In the latter case an additional 0.5 is charged for getting out of the symmetrical structure B-D. This is an empirical question, and we are going to test the predictive power of RhD with both options.

6. Type of rhetorical relation

Some relations make discourse units closer than it is suggested by the RhD measurement (cf. Sanders et al. 1993). In this section we discuss how far the influence of a relation type can go. There are at least two different kinds of instances when such influence seems obvious.

First, Kehler (2002) has convincingly demonstrated that specific types of rhetorical relations impose specific constraints on discourse referents, thus limiting their interpretation scope and allowing

to avoid ambiguities. This is true of Kehler's "resemblance" relations, such as elaboration-additional relation. In such relations, it is normally the topic of the anterior clause that is being elaborated in the following clause. With a high degree of certainty one can predict that it is this topic that will be mentioned and pronominalized.

In the corpus, we could not find natural examples like those in Kehler (2002) (by the way, constructed ones), in which just the relation type alone makes the difference. But the inspection of individual instances of the elaboration-additional relation makes it clear that in such contexts pronominalization is possible at unusually high distances.

In Kibrik (1999) this phenomenon was taken into account by introducing a special factor "predictability": there are contexts in which it is almost inevitable that a referent is mentioned in a certain clause.

The second type of rhetorical relations that may influence antecedent accessibility is those relations that connect embedded clauses. In the canonical version of the RST (Mann and Thompson 1988) embedded clauses were not treated as separate discourse units. However, this decision created serious problems, as it is not infrequent that very usual rhetorical relations hold between two or more clauses embedded into one and the same matrix clause, for example, *Caesar said that he came, saw, and conquered*. Carlson et al. (2003) decided to identify every clause, including even semi-clauses, as separate discourse units. In order to connect such clauses to the rhetorical net, they used some pre-existing relations, such as "purpose" in Fig. 3 above or "circumstance" in Fig. 7 below, and in some cases introduced special relations, such as "attribution" in Fig. 8 below.

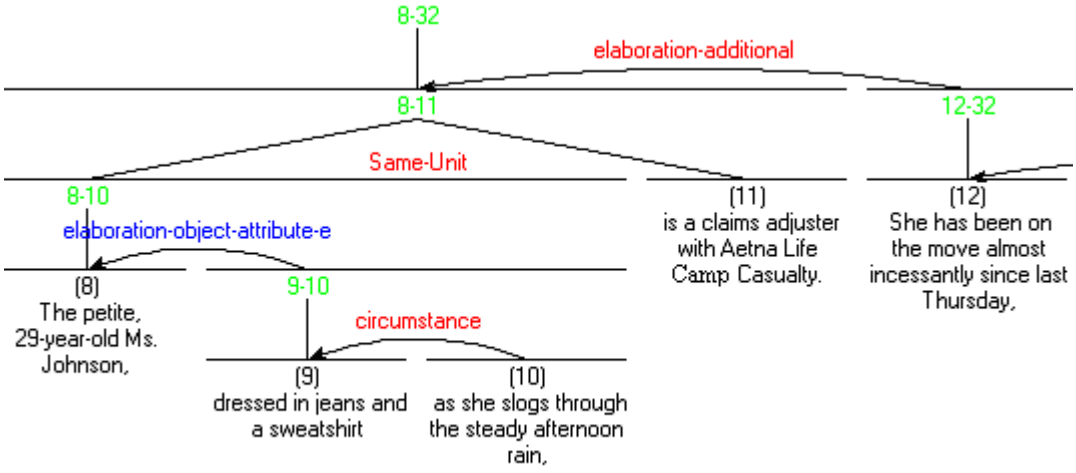

Fig. 7

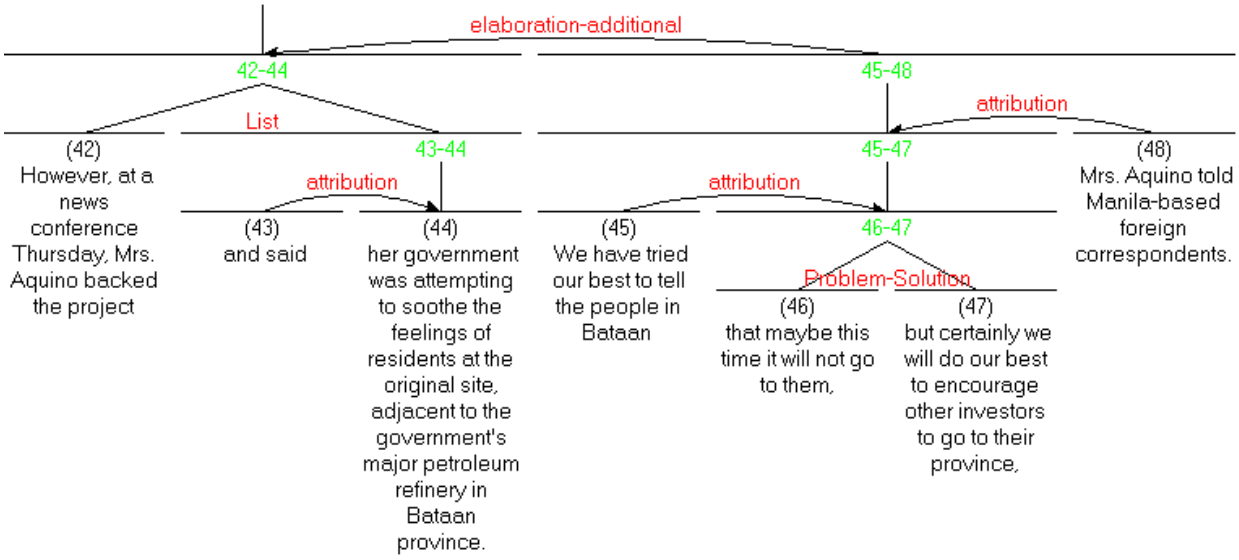

Fig. 8

As a result of these conventions in the RST corpus, syntactically tightly knit clauses (or semi-clauses) count as separate discourse units thus increasing average rhetorical distances greatly. We consider making an adjustment to the rhetorical structure in such instances. We envision two possibilities: either to collapse infinitival and similar clauses with their matrix clauses, thus making the RhD=0, or to compute rhetorical distance between the embedded and the matrix clause as 0.5.

Another problem with the RST corpus is that the relation "attribution" puts the main clause introducing reported speech into the satellite position while the content (reported speech) is the nucleus. see Fig. 8. An artifact of this decision is that the RhD between two adjacent main clauses may turn out quite high in the rhetorical structure – this is what happened to main clauses 42, 43, and 48 in Fig. 8. We are going to make an adjustment in this case too, perhaps reinterpreting the direction of dependency between the nucleus and the satellite, in addition to counting the RhD between the matrix and the embedded clause as 0.5.

In any case, we are going to include coding for the type of rhetorical relations in our database and thus control for this additional factor as potentially affecting referential choice.

7. Multiple antecedents

Because of the non-linear character of rhetorical graphs, there may be more than one candidate for a NP's rhetorical antecedent, consider Fig. 9 below.

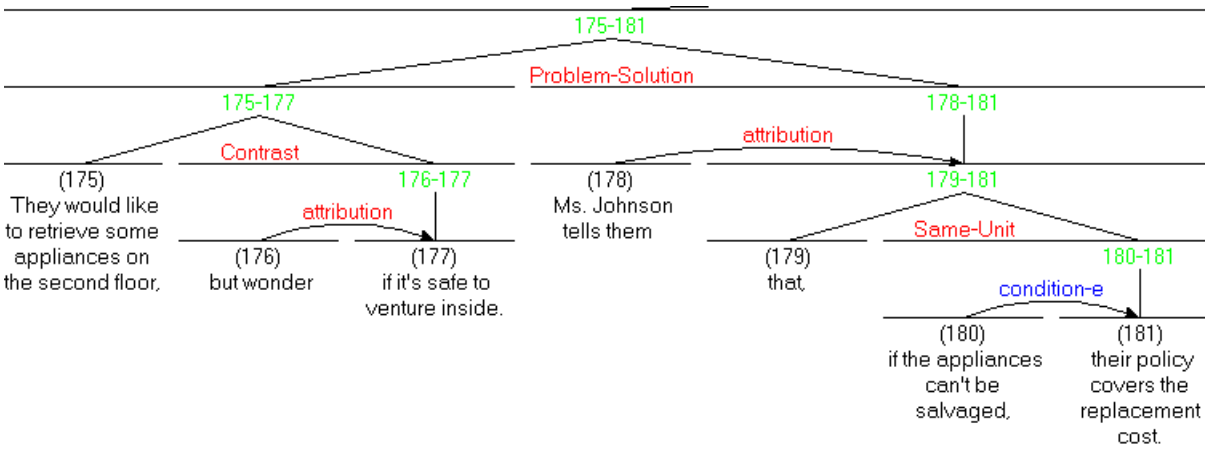

Fig. 9

There is an anaphoric pronoun them in 178, and two previous mentions by the pronoun they in 175 and by zero anaphor in 176. 176 contains the linear antecedent, but which one is the rhetorical antecedent? We propose the following principle:

- choose the antecedent that is the closest in terms of rhetorical distance

According to the conventions discussed above, the RhD from 178 to 176 is 3.5 while the RhD from 178 to 175 is 2.5. Therefore, 175 is chosen as the antecedent clause. This proposal conforms to the principles of the Veins Theory (Cristea et al. 2000).

Of course, one cannot exclude the possibility that occasionally two candidate antecedents would turn out at the same rhetorical distance from the anaphor. It is still important to choose one of them as the actual antecedent, because among the factors affecting referential choice are certain antecedent's properties, such as its syntactic role in its clause. Explorations of possible considerations to take into account in such instances are underway. Among the criteria we consider are:

- choose the candidate antecedent that is linearly closer to the anaphor
- choose the candidate antecedent that has a more prestigious syntactic role
- choose the candidate antecedent that is connected to the anaphor by a tighter rhetorical relation, see Kehler 2002.

8. Conclusion

In this paper we discuss discourse structure as a factor of referential choice. We have explored three specific problems of the interface between rhetorical discourse structure and referential choice: nuclearity status, rhetorical relation type, and competing antecedents. This has been qualitatively investigated on the RST Discourse Treebank data.

Rhetorical distance does not predict a specific referential choice, but is one of its principal factors. A proper rhetorical distance method must clarify probability matters at least; at best, it must incorporate fundamental principles of discourse organization as well. With the help of rhetorical distance, one can set up preferences, according to which the choice of pronouns can be predicted. The understanding of general principles of discourse structure organization together with their interaction with referential choice is essential for the development of theoretical as well as computational models.

References

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In: Current Directions in Discourse and Dialogue. Jan van Kuppevelt and Ronnie Smith eds., Kluwer Academic Publishers.

Cristea, Dan, Nancy Ide and Laurent Romary. 1998. Veins Theory. An Approach to Global Cohesion and Coherence. In Proceedings of Coling/ACL, Montreal, August 1998.

Cristea, Dan, Nancy Ide, Daniel Marcu, and Mihai-Valentin Tablan. 2000. Discourse Structure and Co-Reference: An Empirical Study. In: Proceedings of the 18th International Conference on Computational Linguistics COLING'2000, Luxembourg, July 31-August 4.

Fox, Barbara. 1987. Discourse Structure and Anaphora: Written and Conversational English, Cambridge University Press.

Givón, Talmy (ed.) 1983. Topic continuity in discourse: A quantitative cross-linguistic study. Philadelphia, PA: John Benjamins.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. Computational Linguistics, v.12.

Kehler, Andrew. 2002. Coherence, reference, and the theory of grammar. Stanford, Calif.: CSLI publ.

Kibrik, Andrej. A. 1996. Anaphora in Russian narrative prose: A cognitive account. In: B. Fox (ed.), Studies in Anaphora. Amsterdam: John Benjamins.

Kibrik, Andrej A. 1999. Reference and working memory: Cognitive inferences from discourse observation. In: K. van Hoek, A.A. Kibrik, and L. Noordman (eds.), Discourse Studies in Cognitive Linguistics. Amsterdam and Philadelphia: John Benjamins.

Kibrik, Andrej A. 2002. Discourse types, genre schemata, and rhetorical relations. Paper presented at the 6th Conference on Conceptual Structure, Discourse, and Language. Rice University, Houston, Texas, October 11-14, 2002.

Mann, William C. and Sandra A. Thompson (1988), Rhetorical Structure Theory: Toward a functional theory of text organization, TEXT 8(3).

Sanders, Ted J.M., Wilbert P.M. Spooren, and Leo G.M. Noordman. 1993. Coherence relations in a cognitive theory of discourse representation. Cognitive Linguistics 4(2).

Toole, Janine. 1996. The effect of genre on referential choice. In: Reference and referent accessibility. Ed. T. Fretheim and J. Gundel. Amsterdam: Benjamins.

Wolters, Maria. 2001. Towards Entity Status. PhD Thesis. University of Bonn.