Corpus methods in studies of pronominal anaphora: annotation requirements and methodological strategies

Olga N. Krasavina, Maik Busch

Discourse annotation tasks have been enjoying particular interest in corpus linguistics in the last five-ten years. Given the general complexity of discourse concepts – and coreference is not an exception, no uniform annotation standard for coreference exists to date, and very few coreferentially annotated corpora are available. The current research is based on the RST Discourse Treebank, which we annotated with anaphoric links. The "theoretical" interest of our research are the third person pronouns, considered from a production-oriented perspective We discuss in detail two main challenges we faced in the process of working on this task, namely selection of theoretical approach and annotation methodology. Issues concerning antecedent definition, set of annotation parameters, and exploitation of rhetorical structure annotation are addressed. This work was necessary for preparation of a larger corpus study. Construction of a referentially annotated corpus and of necessary software for its exploitation can be useful both for general understanding of pronominal reference in discourse and for certain NLP tasks, such as anaphor resolution or generation.

1. Introduction

It has been repeatedly argued that the third person pronouns (further – pronouns)[1] are used under high referent's activation in memory of the speaker/addressee. According to our observations, however, there are differences within the class of the pronouns, in conditions under which they are used, including the activation level. Since these differences are very subtle, they need to be accessed empirically on a large data set.

The purpose of our study is twofold. Primarily, it is an empirical research of pronominal reference in discourse. In order to construct a corpus for such purpose, it is necessary to develop an appropriate annotation scheme. The current paper is devoted to the issues related to this second purpose.

The structure of the article is as follows: in Section 2 we describe the annotation of a pilot sample, with particular focus on the annotation methodology and discussion of problem cases. In Section 3 conclusions and directions for the future work are outlined.

2. Problems and proposed decision methods

[1] We exclude the reflexive pronouns from consideration here, since their uses can be basically explained by certain syntactic constraints.

This work is carried out under an assumption that discourse structure plays an important role in referential choice. Mann and Thompson (1988) provide a solid explanatory model of discourse structure, namely Rhetorical Structure Theory (RST).

The corpus required for the purposes of this study should meet the following basic requirements. It should: 1) contain coreference annotation; 2) contain discourse structure annotation; 3) contain at least several thousand pronouns.

Annotation for rhetorical structure is a costly process: it can be accomplished only by well-trained individuals, and demands much time and concentration. Therefore, it is more reasonable to add coreference annotation to the existing rhetorical structure annotation than vice versa. So, we chose the RST Discourse Treebank as the basis of our investigation, with the intention of adding coreferential links to its RST mark-up. Except for the Potsdam Commentary Corpus, which is under construction at the present moment, RST Discourse Treebank appears to be the only existing corpus with this sort of discourse annotation.

The RST Treebank consists of 385 Wall Street Journal articles from the Penn Treebank, or 176,000 words, with 3762 pronouns. According to the RST (Mann and Thompson, 1988), a text can be split into *elementary discourse units* (EDUs), which often coincide with a clause. The discussion of criteria, according to which something is considered to be an EDU is beyond the scope of this paper, but we believe the criteria used for annotation of RST Discourse Treebank are of high reliability (Carlson and Marcu, 1999). EDUs, as well as units consisting of more than one elementary unit are interconnected by means of certain relations, called *rhetorical relations*. In RST Discourse Treebank, 78 relations are annotated: 53 of these are asymmetric, or mononuclear and 25 are symmetric, or multinuclear. The corpus is accompanied by the RST annotation tool – a program used for annotation and visualization of rhetorical trees. In Figure 1, an example of a rhetorical tree is presented, which is a part of a bigger rhetorical tree. Pointed arrows indicate asymmetric relations and follow from satellite to nuclear EDUs.

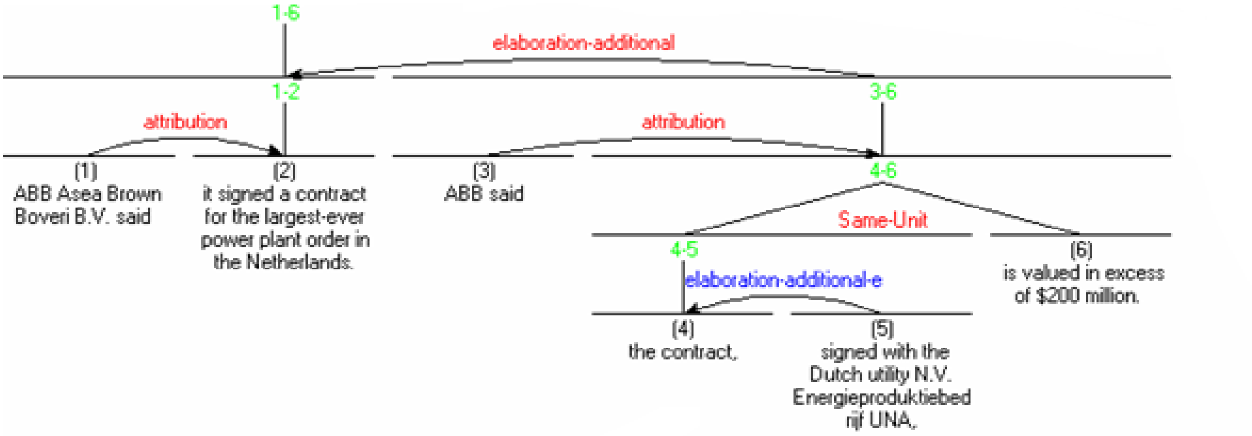
1-6
elaboration-additional
1-2
3-6
attribution
attribution
(1)
ABB Asea Brown Boveri B.V. said
(2)
it signed a contract for the largest-ever power plant order in the Netherlands.
(3)
ABB said
4-6
Same-Unit
4-5
elaboration-additional-e
(6)
is valued in excess of $200 million.
(4)
the contract,
(5)
signed with the Dutch utility N.V. Energieproduktiebedrijf UNA,

Figure 1. Excerpt from a rhetorical graph. Example 1.

2.1 Coreference annotation

Most of the existing work in coreference[2] annotation has been devoted to applied tasks such as anaphor resolution, rather than to theoretical investigation of pronominal reference. The closest NLP domain to our task is anaphora generation, within which rather sophisticated annotation schemes have been suggested (e.g. Tutin et al., 2000). After a careful analysis of existing frameworks, we came to the conclusion that none of the existing annotation schemes could be

[2] Coreference is a relationship between two entities which refer to the same entity in the discourse (Haliday and Hasan, 1976).

adopted completely for our study, although this experience was indeed very helpful. In order to develop an annotation scheme for our study, initially a sample from the RST Discourse Treebank containing 150 pronoun occurrences was annotated with anaphoric[3] links.

The annotation was accomplished semi-automatically, with the use of PAlinkA - the tool developed by Constantin Orasan at the University of Wolverhamton (Orasan, 2003). We chose PAlinkA because it is a platform independent, stable and extendable tool, which enables adding coreferential links to files already containing other types of annotation and which is simple to use. During an annotation process, one can switch between manual and semiautomatic annotation options. Both input and output format is XML.

In order meet the PAlinkA input requirements, the RST Treebank documents were first saved in SGML format, by means of a special option of the RST Tool, and next converted in PAlinkA-compatible XML format by means of the Perl-script. Finally, the files were tokenized, that is, split into words. In Figure 2, an excerpt of a resulting XML-file is presented. Element "node" with its attributes codes information as to the rhetorical structure: "PARENT" is a nuclear node, "RELNAME" is a relation of this node to the node it is attached to. If the value of "RELNAME" is 'span', the parent node is a group of nodes (see Figure 1 for the visualization of the corresponding rhetorical graph). "EXP" stands for markables. "REF SRC" points to the previous mention of a referent in a chain. Chain members, their sequence and total number were later extracted by means of an XQuiry script, written specially for this analysis.

```
<nodeID="1"PARENT="2"RELNAME="attribution"><EXPID="1"><W>ABB</W
><W>Asea</W>>Brown</W><W>Boveri</W><W>B.V.</W></EXP><W>said</
W></node><nodeID="2"PARENT="5001"RELNAME="span"><EXPID="2"><RE
FSRC="1"/><W>it</W></EXP><W>signed</W><W>a</W><W>contract</W>W>
for</W><W>the</W><W>largestever</W><W>power</W><W>plant</W><W>or
der</W><W>in</W><W>the</W><W>Netherlands</W> <W>.</W> </node>
```

Figure 2. Example of an XML format.

2.2 Analysis parameters

The basic goal of studies in anaphor resolution is to develop the smallest and cheapest set of parameters, according to which the inappropriate candidates for antecedents are sorted out. Our goal is, however, to consider as many potential factors as possible. In the future, this set will be reduced to those factors that will turn out to be the most relevant ones. In this study, examples containing pronouns were extracted automatically to an external database, where the values of the factors were filled in for each example, partly automatically, partly per hand. As a whole, we included 30 parameters, which can be subdivided into four main groups: **anaphor/antecedent features** (e.g. grammatical form and function), **referent[4] features** (e.g. semantic class) and **discourse-level features** (e.g. linear distance from anaphor to antecedent). Uses of *it* which can be explained syntactically were not excluded, since this data may be of general interest. Pronoun occurrences in titles and quotations were not excluded either, still with an aim to investigate these cases separately from the others. RST Treebank annotation distinguishes between titles and the rest of the text.

Our hypothesis is that the information as to whether the referent is *a protagonist*, *an animate entity* or whether a current referent mention has *a topical status* is relevant. We distinguish between protagonist and topic, because protagonist can denote only animate entities and it

[3] By "anaphoric", we mean phrases that refer back to a previously mentioned entity.

[4] By "referent", we mean an entity to which the reference is carried out within the current document.

normally does not change throughout the whole discourse, while topic can be used for both animate and inanimate entities and can change many times throughout the discourse.

We pay special attention to a parameter of rhetorical distance derived from the RST (Mann and Thompson, 1988). Although relatively little studied as of today, it appears to be very important. So, in Fox (1987) and Kibrik (2000), it was shown that rhetorical distance often has a more powerful explanatory force than linear and syntactic distance measurements[5]. This parameter is explained in more detail in 2.6.

2.3 Antecedents of referring expressions

Most difficulties and unresolved problems were encountered when determining the antecedent of a referring expression. Some of these problems are addressed in this section.

2.3.1 Linear and rhetorical antecedents

A linear antecedent is understood as the closest previous mention of a referent according to a linear structure of the text, a rhetorical antecedent – according to rhetorical structure. In Kibrik (2000), syntactic roles of linear antecedent and of rhetorical antecedent are distinguished, both factors being influential in referential choice of pronouns and full NPs. It may also be important to distinguish between linear and rhetorical antecedents when calculating rhetorical distance.

In our pilot sample, we annotated the *linearly* previous mention of the referent as an antecedent, for the sake of simplicity. There are a number of problems connected with determining the rhetorical antecedent. For example, what must be considered a rhetorical antecedent, when a satellite precedes the nuclear EDU, as [1] and [2] in Figure 1? Or consider the case where the nuclear-satellite order is "direct" (i.e. nuclear precedes the satellite), but the referent is mentioned twice in the nuclear predication: once by means of a full NP, once by means of a possessive pronoun, as indicated in Figure 3. The question is: what is the antecedent of *he* in (9): *Judge Thomas M. Jenkins* or *his*?

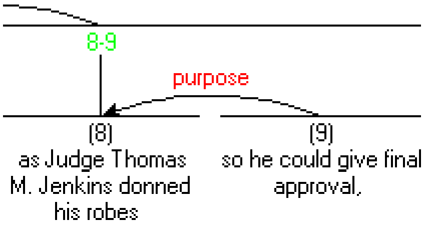
8-9
purpose
[8]
as Judge Thomas M. Jenkins donned his robes
[9]
so he could give final approval,

Figure 3. Excerpt from a rhetorical graph. Example 2.

2.3.2 Length of the antecedent

The question of the length of the antecedent is often addressed in works on anaphor resolution, with a purpose of determining the "minimal strings" an anaphor resolution program is able to recognize. To make an appropriate decision, one needs to consider two possible values of this parameter and choose the weightiest one: value that is most relevant for modeling the referential choice or the one enabling the automatic antecedent recognition and, consequently, automatic coreference annotation. The former is likely to correspond to the maximum length of the antecedent (i.e. NP plus all its dependents), while the latter – to the minimal length (word/group of words which is/are expected to be used for referent nomination most frequently).

Another question that occurs in respect to the length of the antecedent is as follows. The restriction of RST-annotated files is that rhetorical relations within EDUs remain to be unaccounted for. For example, does *it* in (1) refer to *Shiseido Co* or to *Shiseido Co., Japan's leading cosmetics producer*?

[5] The difference between linear and syntactic distance is that the former is normally measured in EDUs, while the latter – in sentence boundaries.

(1) Shiseido Co., Japan's leading cosmetics producer, said it…

Following the logic of the RST, one can break the sentence presented in (1) into two sentences, which are connected by a rhetorical relation of elaboration: 1) *Shiseido Co. is Japan's leading cosmetics producer.* 2) *Shiseido Co. said ...*

As an annotation solution for such cases, we applied the RST distinction between restrictive and non-restrictive attributes. In (1), the attribute is non-restrictive, so we annotated *Shiseido Co.* and *Japan's leading cosmetics producer* as two separate markables, *it* linked to the latter. By doing so, we do not mean that these markables are connected by coreference relation, but take the actual discourse structure into consideration. The optimal decision would be to indicate such a type of relations in a specific way.

In (2), *it* refers to the whole previous EDU (borders of EDUs are marked by square brackets). Since the attribute introduced by means of a *that*-clause is restrictive, we marked the whole node as an antecedent.

(2)[**Energetic and concrete action has been taken in Colombia during the past 60 days against the mafiosi of the drug trade**], [but **it** has not been sufficiently effective].

In case of non-nominal antecedents, it often happens that an antecedent goes beyond one EDU. Here our annotation options were limited by a PalinkA restriction that we discovered during the annotation process: such strings cannot be annotated by PalinkA as markables.

It is often the case that pronouns are used after zero-mentions of a referent.[6] In our sample, we annotated zero mentions as well.

2.3.3 Multiple referents

PAlinkA provides an option for marking multiple antecedents, see Figure 4 for XML representation (the corresponding IDs are marked bold):

```
<EXP ID="54"> <REF SRC="53"/> <W>William</W> </EXP> <W>and</W>
<EXP ID="55"> <REF SRC="53"/> <W>Margie</W> <W>Hammack</W> <EXP
ID="56"> <REF SRC="55"/><W>Mrs.</W>
<W>Hammack</W></EXP> </EXP> <EXP ID="59"> <REF SRC="54 56"/>
<W>The</W> <W>Hammacks</W> <W>&apos[7];</W> /EXP><W>own</W>
<W>home</W>
```

Figure 4. Multiple referents in a PAlinkA-annotated file.

However, it is not clear how to proceed if the linear sequence in the coreference chain is as follows: *Mr. and Ms.Hammack – the Hammacks –they – Mr.Hammack,* (in case all these referents are marked as coreferential). Can *they* or *the Hammacks* be really considered to be an antecedent for *Mr.Hammack*? Or should the last explicit mention of Mr. Hammack be marked as an antecedent? The latter would not contradict our antecedent definition, but it is unlikely that the addressee would retrieve the representation of *Mr. Hammack* rather then that of *the Hammacks* which was recently mentioned. The alternative solution would be to define this relation as "part-whole", which is discussed in more detail in the next subsection.

2.4 Relation types between anaphor and antecedent

Although the identity relation appears to be the most frequent one in case of pronouns, other relation types exist. For example, the following referents stand in "whole-part" relation to each other: *a unit of DPC Acquisition Partners - DPC Acquisition*. Further in the same text we see another example of "part-whole" relation, or the subset of this, which can be termed as

[6] By "zero antecedents" we mean syntactic zeros, that is, zeros occurring as a result of such procedures as ellipsis.

[7] The sign "&apos" denotes apostrophe in XML.

"member-organization": *Dataproducts officials - they - Dataproducts board -it - Dataproducts – it.*

We annotated *Dataproducts Corp*. and *Dataproducts officials* as identical, although it is clear that they are not: *Dataproducts* is a company, not an animate entity, and it cannot decline anything. It is an interesting phenomenon, by which the properties of a member or members are projected on the organization. This is frequently observed in financial reports, particularly in constructions like 'X said', where X is either an organization, or its 'speaker', 'official', etc. It is essential to find another annotation method for this relation, rather than marking *Dataproducts officials – they – Dataproducts board – it – Dataproducts – it* as one coreference chain.

2.5 Relation types between the components of the possessive NP

We annotated only the possessive determiner of the NP (for situations like X - X´s Y), not the whole NP (see Figure 4). However, it may be useful to consider the relations between the elements of a possessive NP. Different classifications describe these relations with varying degrees of granularity. On the basis of our material, we developed a typology with a medium detail of elaboration, which should be included in the final annotation scheme: "X has an inherent part Y"; "X has an associated part Y"; "X has a social relation Y"; "X is a member of Y"; "X is a performer of Y"; "X is an owner of Y".

2.6 Methods of calculating rhetorical distance

Already in early cognitive-psychological accounts (e.g. Givon, 1983), the distance between the referent mentions was argued to be one of the weightiest factors of referential choice: pronouns tend to be used close to the antecedent. Rhetorical distance is a powerful measurement of how close the anaphor and the antecedent really are, since it grasps the relation types between discourse units, which is not always explicitly reflected in the syntactic structure of the discourse.

At the moment, no uniform methods of rhetorical distance computation exist. Simply defined, rhetorical distance is counted in the same manner as syntactic (of linear) distance, that is, in the number of steps to the left one needs to make along the graph, in order to reach the antecedent, with the graph as the only difference – rhetorical or syntactic. Linear and rhetorical distance values can coincide, as in a simple example, indicated in Figure 1. Here, rhetorical distance between *it* and its antecedent is 1 – the same as the linear distance.

This measurement appears to be useful only in two cases: at short syntactic distances, if it allows considering the important distinctions which are not reflected in the syntactic structure, and at long distances, if it allows neglecting the irrelevant syntactic distinctions.[8] With respect to the complexity of rhetorical structure in some cases, we decided to implement several methods and evaluate their efficiency on the basis of the results. An algorithm was developed, which involved computation of six parameters:

- number of nuclear nodes of symmetrical relations (here and below, parameter is counted separately for both target nodes);
- number of satellite nodes of asymmetrical relations;
- number of satellite nodes of asymmetrical relations.

Eventually, the rhetorical distance definition is supposed to be derived from one of these parameters, probably in combination with another one (or ones). We do not exclude the

[8] We do not cite the examples here, since they demand much space. For some examples, see Kibrik 2000.

possibility that the final method may vary as to genre and linear remoteness of referent mentions from each other.

3. Future work

In this paper, the preliminary work on the pronoun-oriented coreference annotation of RST Discourse Treebank was reported. There are a number of decisions that are still to be made before we can proceed to the annotation of the whole corpus, primarily concerning antecedent determination and relations between the referent mentions. On the basis of the coreferentially annotated pilot sample, a method of calculating rhetorical distance is to be developed, and the set of most significant factors which have to be included into the annotation scheme will be determined.

References

1. Carlson, Lynn and Daniel Marcu. Instructions for Manually Annotating the Discourse Structure of Texts, 1999. URL: http://www.isi.edu/~marcu/ software/ manual.ps.gz
1. Fox, Barbara. Discourse Structure and Anaphora. Cambridge: Cambridge University Press, 1987.
2. Givyn, Talmy (ed.). Topic Continuity in Discourse. A Quantified Cross-language Study. Amsterdam and Philadelphia: John Benjamins, 1983.
3. Halliday, Michael and Ruqaiya Hasan. Cohesion in English. Longman Group Ltd, 1976.
4. Kibrik, Andrej A. A Cognitive Calculative Approach towards Discourse Anaphora. In Proceedings of the Discourse Anaphora and Anaphor Resolution Conference (DAARC 2000). Baker/ Hardie/ McEnery/ Siewierska (eds.), Lancaster, UK: University Centre for Computer Corpus Research on Language, 2000.
5. Mann, William C. and Sandra A Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. In: Text 8 (3), 1988. Pages 243-281.
6. Orasan, Constantin. PALinkA: A highly customisable tool for discourse annotation. In Proceedings of Fourth SIGdial Workshop on Discourse and Dialogue, July 2003..
7. Tutin, Agnes, Francois Trouilleux, Catherine Clouzot, Eric Gaussier, Annie Zaenen, Stephanie Rayot and Georges Antoniadis. Annotating a Large Corpus with Anaphoric Links. In Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000), pages 28-38, Lancaster, UK, November.

URLs

8. PalinkA URL: http://clg.wlv.ac.uk/projects/PALinkA/
9. Potsdam Commentary Corpus. URL: http://www.ling.uni-potsdam.de/cl/cl/res/forsch_pcc.html#D00
10. RST annotation tool. URL: http://www.isi.edu/~marcu/