# **C-3: Coherence and Coreference Corpus**

Cristina Nicolae, Gabriel Nicolae, Kirk Roberts

Human Language Technology Research Institute Department of Computer Science University of Texas at Dallas Richardson, TX 75083-0688 E-mail: {cristina, gabriel, kirk}@hlt.utdallas.edu

#### Abstract

The phenomenon of coreference, covering entities, their mentions and their properties, is intricately linked to the phenomenon of coherence, covering the structure of rhetorical relations in a discourse. A text corpus that has both phenomena annotated can be used to test hypotheses about their interrelation or to detect other phenomena. We present the process by which C-3, a new corpus, was obtained by annotating the Discourse GraphBank coherence corpus with entity and mention information. The annotation followed a set of ACE guidelines adapted to favour coreference and to include entities of unknown types in the annotation. Together with the corpus we offer a new annotation tool specifically designed to annotate entity and mention information within a simple and functional graphical interface that combines the "best of all worlds" from available annotation tools. The potential usefulness of C-3 is discussed, as well as an application in which the corpus proved to be a valuable resource.

## 1. Introduction

As defined by (ACE, 2004), an entity is an object or set of objects in the world, while a mention is a textual reference to an entity. All mentions that refer to an entity are said to corefer (or be coreferent) with each other. Discourse relations (rhetorical relations that hold between segments of a natural language discourse) are defined in scientific literature (Mann & Thompson, 1998; Hobbs, 1985; Wolf & al., 2003) based on the entities or situations involving them that can be inferred from the segments considered. The phenomenon of coreference, which covers knowledge about entities, their mentions and their properties, is intricately connected with the phenomenon of coherence, which covers discourse relations. (Hobbs, 1979) explains this connection by using knowledge of discourse coherence to solve coreference, while assuming the existence of a real world knowledge base that is still far from reach. On the other hand, knowledge of coreference is more easily available and detectable and can be of great help in detecting coherence, i.e., the rhetorical relation structure of a discourse.

To our knowledge, there is no existing text resource aside from ours to have both discourse coherence and coreference annotations. The most widely known coreference corpora are the ones offered by the Message Understanding Conference, e.g., (MUC-6, 1995) and the Linguistic Data Consortium, e.g., (ACE, 2004), but apart from coreference information both sets of corpora are annotated with relations between entities, not between discourse segments. The most widely known coherence corpora are Discourse GraphBank (Wolf & al., 2003), RST Treebank (Carlson & al., 2002) and Penn Discourse Treebank (Prasad & al., 2008), none of which was annotated with coreference information before the project reported here.

Since it is more accessible to annotators with no advanced linguistic knowledge to annotate coreference rather than

coherence, a coherence corpus was chosen to be annotated with coreference information. Among the three coherence corpora mentioned above, Discourse GraphBank is the one most amenable to annotation because of its volume (much smaller compared to Penn Discourse Treebank), its inclusion of long-distance relations (which are barely annotated in Penn Discourse Treebank), and its superior way of representing discourse structure as a graph rather than a tree (when compared to RST Treebank). Discourse GraphBank was also designed to be easily extensible by keeping the text files separate from the annotation files. For all these reasons, Discourse GraphBank was the coherence corpus selected as the base for the coreference annotation project presented in this paper. The project could also have aimed to annotate a section of a larger corpus, but this choice was made for completeness' sake. The result is a coherence and coreference corpus named C-3<sup>1</sup>.

The rest of this paper is structured as follows. Section 2 offers a look at Discourse GraphBank. Section 3 describes the annotation project. Section 4 includes a discussion of the potential benefits of the new corpus, while Section 5 draws the conclusions.

### 2. A Brief Look at Discourse GraphBank

Discourse GraphBank (Wolf & al., 2003) is a database of 135 texts annotated with discourse relation information. Its text documents are collected from various public sources (such as news items and aptitude tests) and cover a wide range of topics. Two annotators identified in these texts a set of relations taken mostly from (Hobbs, 1985) and (Kehler, 2002). The coherence relations labeled on Discourse GraphBank occur between two discourse segments (units or groups of units) and can be symmetrical or asymmetrical. In symmetrical relations both sides have equal importance, while in asymmetrical

<sup>&</sup>lt;sup>1</sup> Available at http://www.hlt.utdallas.edu/~cristina/c-3.

relations the two sides are called "nucleus" and "satellite", with the nucleus having a higher importance in the document than the satellite. These elements were first defined by William Mann and Sandra Thompson in the seminal paper that presented Rhetorical Structure Theory (Mann & Thompson, 1988). Wolf & al. annotated the following discourse relations on Discourse GraphBank:

1. Resemblance relations: *parallelism, contrast, exemplification, generalization and elaboration.* These relations are identified by inferring two sets of entities from the two discourse segments involved, and then inferring comparisons between members of the two sets. For instance, two discourse segments are in a contrast relation if one or more entities inferred from the first segment are in contrast with one or more entities inferred from the second segment.

2. Causal relations: *explanation (cause/effect), violated expectation and condition.* These relations are identified by inferring a causal connection between the two discourse segments. For instance, two discourse segments are in a condition relation if an event described in the nucleus is conditioned on an event described in the satellite.

3. Other relations: *temporal sequence and attribution*. Two segments are in a temporal sequence if events described in them are in a temporal sequence. Two segments are in an attribution relation if the satellite attributes entities or events described in the nucleus to a source.

# 3. C-3 Annotation Project

## 3.1 Annotated Elements

For each mention, the elements annotated were its extent, its head (or heads) and its properties (mention type, role and metonymic type). An example of a mention is "*the Czechoslovak border*", referring to the real-life entity (of type location) that used to be the border between Czechoslovakia and its neighbors. The extent of the mention consists of its entire nominal phrase ([*the Czechoslovak border*]), while the head of the mention is its representative word (*border*).

Entities were annotated by grouping together the mentions that referred to them. These groups were further annotated with entity properties (entity type, subtype and class). For each document, we annotated non-single mentions (mentions that corefer with others) and single mentions that refer to an entity of one of the seven types labeled by ACE (person, organization, location, facility, weapon, vehicle, and geo-political entity). The reasoning behind this choice is that we wanted to obtain a corpus that offers complete information about the noun phrases involved in coreference and the ones having known entity types, without cluttering the texts by annotating every single noun phrase. "Cluttering" here refers not to the format of the annotations, which are machine-readable, but to the number of annotated elements. Since the scope of this project is coreference, i.e., identifying mentions to entities, the noun phrases that are not connected to others

by coreference and at the same time do not refer to a known entity type represent clutter.

Following are the entity and mention properties annotated in the C-3 project. They respect the guidelines set by the Automatic Content Extraction (ACE) Entity Detection and Tracking (EDT) task defined in (ACE, 2004), with some modifications, which are noted in the following subsection. All the examples are taken from the coreference-annotated Discourse GraphBank files and the definitions are summarized from the original ACE guidelines. The mentions' extents are encased in square brackets and their heads are underlined.

### **Entity Type and Subtype**

1. **PER** (Person): A distinct person or a set of people, fictional or real.

"[<u>Students</u>] can see how the factory of the future operates."

2. **ORG** (Organization - government, commercial, educational, nonprofit, other): Organizations are groups of people defined by an established organizational structure.

"according to [the <u>university</u>]'s model" (educational)

*"[The funding <u>committee]</u> is a non-profit, non-partisan coalition of groups"* (nonprofit)

3. LOC (Location - address, boundary, celestial, water-body, land-region-natural, region-local, region-subnational, region-national, region-international, other): Locations are geographical entities such as geographical areas, water bodies and geological formations.

"The incident occurred well off the Libyan coast in [international <u>airspace</u>]." (celestial)

"The giant plumes blasted into the sky by [volcanoes] may look like ordinary clouds" (land-region-natural)

4. **FAC** (Facility - plant, building, subarea-building, bounded-area, conduit, path, barrier, other): Facilities are buildings and other man-made structures.

"Warren G. Harding died there, in [<u>room</u> 8064] on Aug. 2, 1923" (subarea-building)

"who lives 400 yards from [the airport's main <u>runway</u>]" (path)

5. **GPE** (Geo-political entity - continent, nation, state-or-province, county-or-district, city-or-town, other): GPEs are geographical regions defined by political and/or social groups.

"[<u>West Germany</u>] has more traffic volume than any nation in [<u>Europe</u>]" (nation, continent)

"about 30 miles northeast of [Rome]" (city-or-town)

6. **VEH** (Vehicle - air, land, water, subarea-vehicle, other): Vehicles are physical devices primarily designed to move an object from one location to another.

*"[the jets], assigned to [the aircraft carrier <u>USS John F.</u> <u>Kennedy]</u>" (air, water)* 

*"[The <u>cockpit</u>] with a large chunk of fuselage"* (subarea-vehicle)

7. **WEA** (Weapon - blunt, exploding, sharp, chemical, biological, shooting, projectile, nuclear, other): Weapons are physical devices primarily used to physically harm

living beings or destroy constructions.

*"But my son had no weapon, only [a <u>machete]</u>" (sharp) <i>"elimination of [[a proposed single-warhead intercontinental ballistic <u>missile]</u>, [the <u>Midgetman]]</u>" (projectile)* 

8. **OTH** (Other): All other entities that don't fall under the seven named types above, but only taggable if they are referred to by more than one mention.

"both sides in [a 9-year-old civil <u>war</u>] grow frustrated ... [The <u>war</u>] has claimed an estimated 65,000 lives" "and said it will introduce [a new incentive <u>plan</u>] for advertisers. ... [The new ad <u>plan</u> from Newsweek] ..."

### **Entity Class**

1. **NEG** (Negatively quantified): The entity is quantified to refer to the empty set of the type of object mentioned.

"He said army records show [no <u>troops</u> near Piedra Luna on Nov. 12]"

2. **SPC** (Specific referential): The entity is a particular, unique object or set of objects (even if its name or location are not known).

"discontinued operations in [the third <u>quarter</u>] because of [the planned <u>sale</u>]"

3. **GEN** (Generic referential): The entity is not particular or unique, but represents a type of objects.

"drivers miffed at having to change lanes to get by [a slower <u>car</u>]"

4. **USP** (Underspecified referential): Neither generic nor specific reference; the entity referenced cannot be verified.

"[Fatal accidents – about 8,000 last year]"

5. **ATR** (Attributive/non-referential): The entity is not being used to refer, but to attribute a property to another entity.

"Bernt Carlsson was [U.N. Commissioner for Namibia]"

### **Mention Type**

1. NAM (Names): Proper nouns and nicknames.

"[*Libya*] says the plant in question produces pharmaceuticals."

2. **NOM** (Quantified nominal constructions): Nouns quantified with a determiner, a quantifier or a possessive. *"marked his 1,000th day in captivity with [a vigil outside city hall]"* 

3. **BAR** (Bare nominal mentions): Unquantified nominal constructions.

"investors continue to pour cash into [money funds]"

4. **PRO**: Pronouns except wh-question words and the specifier 'that'.

"who was trapped in [her] bedroom"

5. WHQ: Wh-question words and the specifier 'that'.

"a noted constitutional lawyer [<u>who</u>] is also defending convicted murderer John Joseph Jindler"

6. **PRE** (Premodifier mentions): Mentions that occur in a modifying position before other words.

"[Soviet]-made products for [Soviet] consumers"

7. **HLS** (Headless mentions): Constructions in which the head is not specifically expressed.

"the very angry veterans who are quite abusive on the

phone, [the very <u>courteous</u>] who are pleased with the information"

8. **PTV** (Partitive constructions): Constructions that refer to a part of a mention.

"[<u>one</u> of the predecessors the president most admires]" 9. **CMC** (Conjoined mention constructions)

Constructions that consist of two or more mentions.

*"require them to cease the widespread practice of using* [headlights and blinkers] to pressure slower cars"

10. **APP** (Appositive constructions): Constructions consisting of two or more mentions that refer to the same entity.

"[Launius' wife, Susan]"

11. **ARC** (Complex appositive constructions): Appositive constructions which contain at least a relative clause. *"[Nash, 59, whose real name is Adel Nasrallah]"* 

#### **Mention GPE Role**

This property refers to mentions of GPEs.

1. **GPE.PER**: Mentions of the population of a GPE.

"[<u>Americans</u>] today spend \$15,000 like pocket change"

2. **GPE.ORG**: Mentions of the governing body of a GPE. "[U.S. <u>government</u>] officials said the Navy jets were conducting routine operations"

3. **GPE.LOC**: Mentions of the territory or geographic position of a GPE.

"it could clear the way for Soviet bonds to be sold in [the U.S.]"

4. **GPE.GPE**: Mentions for which no role stands out in the context.

"In 1941, the Nazis attacked [the Soviet Union]."

Mentions marked with a GPE role represent one of the four aspects of a geo-political entity: a population, a government, a physical location and a nation (or state, city, etc.). Mentions that are not connected with a geo-political entity are marked with the equivalent entity type: PER, ORG, LOC.

### Mention Metonymic Type

1. Metonymic: The mention to an entity is used to refer to another entity or entities related to it.

*"the major issue dividing the parties will be whether they speak primarily for [Beijing] or Hong Kong"* 

2. Non-metonymic: The mention has a straightforward reference to an entity.

"[*China*] appears to have reluctantly dropped [*its*] opposition to such activity"

## 3.2 Differences from ACE Guidelines

We slightly adapted the ACE annotation guidelines to prefer referential to attributive interpretations and mark coreference wherever possible. The changes are meant to supplement the narrow scope of the original guidelines that limit identification to seven entity types and disregard the anaphoricity of some mentions, such as premodifiers that modify people.

1. An eighth entity type, OTHER, was introduced to label all entities that do not have one of the seven known types. Mentions were tagged to refer to OTHER entities only if they coreferred with at least one other mention. This way the text is not cluttered by the annotation of every noun phrase, but all the coreferential noun phrases are annotated, which constitutes a wealth of new information. Computers, for instance, are entities not covered under the original ACE guidelines, but are marked in C-3:

"But [<u>Apple II</u>] was a major advance from Apple I"

"In addition, [the Apple II] was an affordable \$1,298."

2. Premodifier mentions (PRE) were considered referential wherever possible, not just attributive – even when they modified people with titles or professions. Premodifiers were only annotated if they either coreferred with other mentions in the text or they referred to an entity labeled with a known type. Annotating all premodifiers would have cluttered the text unnecessarily, but not annotating some of them as referential would have lost coreference information. Here is an example of coreferential premodifiers in C-3:

#### "said [White House] spokesman Roman Popadiuk." "the [White House] Situation Room"

Considering premodifiers to be referential makes it easier, in this case, to capture connections between discourse segments that focus on an organization and segments that focus on spokespeople acting on behalf of the same organization.

3. For ease of use, appositive constructions (APP) are headed by the head of their first clause. All the mentions in the APP are also annotated as referring to the same entity as the APP. This contributes to the consistent property of C-3 that no mention goes without at least one marked head. Example: "[Kathy Drake, a spokeswoman for the Department of Corrections in Atlanta]".

4. Appositive constructions are annotated as ARC if they contain at least one relative clause, even if they consist of only two clauses. This choice was made to differentiate more clearly between "simple" appositions and relative constructions. An example of an ARC that appears in C-3 and would not be marked as such under the ACE guidelines is "*[the injured officer, who was not identified]*".

5. We marked all the heads in conjoined mention constructions (CMC), i.e., we introduced mentions that have more than one head; at the same time, we eliminated the mention type MWH (multiple-word heads). The MWH mentions in the ACE guidelines and the CMC mentions are treated consistently in C-3 by marking all the heads separately. The components of the CMC and the whole structure are tagged only if needed, that is if they corefer with other mentions or refer to entities labeled with known types. This extension to multiple heads was done in order to capture the equality of elements in conjoined constructions. For example, a mention that would have been marked as MWH in the ACE guidelines and is marked as a CMC in C-3 is "[Stena Holding AG and Tiphook PLC]". Coreference is allowed between multiple headed mentions and single headed mentions whenever the single head refers to the set of heads of the CMC, for example a mention coreferent to the one above would be "[two European shipping concerns]".

#### **3.2** Annotation Tool

The 135 files were entirely annotated by one annotator using the gann (Graphical Annotation) Tool<sup>2</sup> developed specifically for an Entity Detection and Tracking annotation task. The best-known annotation tools at the time gann was developed were MMAX2<sup>3</sup> (Müller & Strube, 2006), the LDC ACE Annotation Tool<sup>4</sup> and WordFreak<sup>5</sup>. MMAX2 was built as a flexible multi-level annotation tool able to create coreference chains, but it did not offer the functionality to add attributes (such as type and subtype) to coreference chains to turn them into typed entities, a functionality needed by the C-3 task. The LDC ACE Annotation Tool was built specifically for ACE annotations, but was unable to select multiple heads for a mention, a part of the C-3 task. WordFreak, a plugin-based tool, offered the potential of adding desired functionality by writing new plugins. However, the ACE plugin that was available with it had multiple limitations: it did not mark entity subtypes, only one head could be selected for mentions, and the file format did not intuitively capture the entity/mention group structureall mentions referring to the same entity shared the same entity id, while entities did not appear as elements. The latter limitation was also present in the MMAX2 file format. Most, if not all, of these problems could have been solved by rewriting or correcting the plugin used for the annotation and by running postprocessing scripts on the data, but after considering all compromises we chose to start fresh with a new tool that combined the best of all worlds from the perspective of our task.

*gann* offers a graphical interface similar to WordFreak's, designed to be simple yet functional. This interface allows a user annotating a plain text document to select mention extents and heads. The selection is done at the word level, not the character level, which avoids unwanted word-cutting. Colors are used to highlight currently selected mentions and chains. The selection area is illustrated in Figure 1.

The <u>U.S.</u> and <u>Soviet Union</u> are holding technical <u>talks</u> about possible repayment by <u>Moscow</u> of \$188 million in pre-Communist Russian <u>debts</u> owed to the <u>U.S. government</u> .
<u>the State Department</u> said.
If the <u>debts</u> are repaid,
it could clear the way for <u>Soviet</u> bonds to be sold in the <u>U.S.</u>
However,
after two meetings with the Soviets,
a <u>State Department spokesman</u> said that
it's "too early to say" whether that will happen.
Coincident with the <u>talks</u> ,
the State Department said
<u>it</u> has permitted a <u>Soviet bank</u> to open a <u>New York branch</u> .
The branch of the Bank for Foreign Economic Affairs was approved last spring
and opened in July.

Figure 1: gann text selection area.

Mentions can be grouped into entities, and basic operations can be performed on entities and mentions (e.g., merge and remove). Each new mention is immediately attached to an existing or a new entity, which

<sup>&</sup>lt;sup>2</sup> Available at http://www.hlt.utdallas.edu/~gabriel/gann.

<sup>&</sup>lt;sup>3</sup> http://mmax2.sourceforge.net

<sup>&</sup>lt;sup>4</sup> http://projects.ldc.upenn.edu/ace/tools/jan\_2004\_tool

<sup>&</sup>lt;sup>5</sup> http://wordfreak.sourceforge.net

makes *gann* a one-step annotation tool. There is no unattached mention at any time in the process, which intuition suggests as natural considering a mention is defined as a reference to an entity. *gann* presents to the user a navigable list of entities and their mentions at any time; selecting a mention or entity on the list selects it in the text and viceversa. The tool also offers the possibility to select multiple heads for each mention. The list of entities and mentions is illustrated in Figure 2.

Figure 2: gann entity and mention list area.

Both entities and mentions can be assigned properties, i.e., types, subtypes, classes and roles (Figure 3). Entity types and subtypes are defined in a hierarchy and the set of all properties is easily customizable.

Finally, the annotation tool is portable (being written in Java) and has no external dependencies.

### 3.3 Annotation Procedure

The annotation process consisted in marking all the elements enumerated in subsection 3.1 on the texts, using the tool described in subsection 3.2. The annotation was performed in a few passes over the corpus with care to apply the experience gained while annotating the entire corpus to earlier annotations, to check and recheck for missed elements and to ensure an overall consistency of the markings.



Figure 3: gann entity and mention properties area.

The process encountered straightforward elements and difficult elements. The easiest items to annotate were simple and intuitive properties like mention type and certain entity types, e.g., FAC and VEH.

Challenging cases covered a gamut of difficulties, such as:

1. Requiring nontrivial historical, geographical and cultural knowledge in order to understand a reference. As an example, the reference to "*the [Hoosier] capital of Indianapolis*" assumes the cultural knowledge that a resident of the State of Indiana, US is also known as a Hoosier.

2. Confusion when choosing between GPE/other and LOC/land-region-natural as entity subtypes. For example, in *"the emperor in whose name the militarists had overrun [East Asia]"*. East Asia is a geologically designated region, but it is unclear whether it functions as a political entity in the context.

3. Ambiguity when deciding whether to label an entity as a WEA. A reference to Agent Orange is a pertinent example of this ambiguity: [the chemical <u>defoliant</u> – sprayed over Southeast Asia during the 1960s by the U.S. military in an attempt to deprive Communist troops of crops and cover] – caused cancer, birth defects in their

*children and other illnesses.*" It is unclear whether Agent Orange can be classified as a chemical weapon, because its primary function is of a herbicide, but it physically harmed living beings by being used as such.

4. Ambiguity between generic and underspecified labels, for example in "a consultant who made money obtaining information for [large defense contractors]". The mention "large defense contractors" refers to a kind of contractors, so it fits the definition of a generic entity, but later this choice is challenged when the contractors that benefitted from the consultant's information are enumerated, making the entity underspecified.

At the time of this paper, C-3 is a work in progress, in that it has not been verified entirely by a second annotator. In order to capture a preliminary view of the inter-annotator agreement, the second annotator was presented with 10 files at random from the corpus and was requested to note agreement and disagreement with the first annotator's choices. We computed the inter-annotator agreement score according to the methodology proposed by (DeCristofaro et al., 1999) for coreference annotation. Agreement was computed on markables and properties, that is, we measured the percentage of mention extents and properties, respectively, that were agreed on by the second annotator. Part of the mention properties were inherited from their entities.

The agreement on properties, Cohen's Kappa agreement coefficient (Cohen, 1960), was computed from the following expression:

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the percent agreement (actual agreement) and P(E) is the expected agreement. Considering a property *p* and *N* mentions in a corpus, P(A) is the fraction of mentions out of *N* for which the two annotators assigned the same value to *p*. If property *p* takes values in the domain *V*,

$$P(E) = \sum_{v \in V} \left( \frac{c_1(v, p) + c_2(v, p)}{2 * N} \right)^2$$

where  $c_i(v, p)$  is the number of times annotator *i* assigned value *v* to property *p*.

The agreement on two sets of markables  $(s_1 \text{ and } s_2)$  is computed from the following expression:

$$Agr(s_1, s_2) = \frac{2*c}{a+b}$$

where  $a=|s_1|$ ,  $b=|s_2|$ , and *c* is the number of mentions marked in  $s_1$  that were marked with exactly the same boundaries and the same heads in  $s_2$ .

Given these expressions, the inter-annotator agreement on properties is reproduced in Table 1. The average agreement on markables is 0.98.

	P(A)	P(E)	k
mention type	0.99	0.18	0.98
metonymic type	0.99	0.97	0.83
entity type	0.94	0.19	0.92
entity class	0.93	0.61	0.82

Table 1: Inter-annotator agreement on mention properties.

The high agreement values are partly due to the fact that the second annotator was not required to re-annotate the files, just to agree or disagree with the existing annotations, so the level of scrutiny was lower.

### 3.4 C-3 Annotation File Format

C-3 continues the Discourse GraphBank rule of keeping the texts separate from the annotation files corresponding to them. The general format of an annotation file saved by *gann* is represented in Figure 3. The mention and entity attributes encased in square brackets may take the

#### following values:

```
[entity-id] := <number>
[mention-id] := <number>
[annotator-id] := <number>
```

[entity-type] := Person | Organization | Location
| Facility | GPE | Vehicle | Weapon | Other

```
for [entity-type] = Organization:
[entity-subtype] := Government | Commercial |
Educational | Non-profit | Other
```

for [entity-type] = Location: [entity-subtype] := Address | Boundary | Celestial | Water\_Body | Land\_Region\_natural | Region\_Local | Region\_Subnational | Region\_National | Region\_International | Other

```
for [entity-type] = Facility:
[entity-subtype] := Plant | Building |
Subarea_building | Bounded_Area | Conduit | Path
| Barrier | Other
```

```
for [entity-type] = GPE:
[entity-subtype] := Continent | Nation |
State-or-Province | County-or-District |
City-or-Town | Other
```

```
for [entity-type] = Vehicle:
[entity-subtype] := Air | Land | Water |
Subarea_vehicle | Other
```

```
for [entity-type] = Weapon:
[entity-subtype] := Blunt | Exploding | Sharp |
Chemical | Biological | Shooting | Projectile |
Nuclear | Other
```

```
[entity-class] := NEG | ATR | SPC | GEN | USP
[mention-type] := NAM | NOM | BAR | PRO | WHQ | PRE
| HLS | PTV | CMC | APP | ARC
[mention-role] := N/A | PER | ORG | LOC | GPE
[metonymic-type] := true | false
[extent-range] := <number>..<number>
[head-ranges] := <number>..<number>,
<number>..<number>, [...]
[comment] := <string>
```

The annotator id is set for each entity and mention to the annotator who last changed its properties; this facilitates computing the inter-annotator agreement. Comments (attached to entities and mentions alike) are not required, but they can come in useful if, for instance, the annotator needs to explain choices or mark types for OTH entities. As an example, here is an entity from the annotation file generated for file 8 of the corpus. The entity is a commercial organization (bank) that is mentioned three times in the file, by two nominal mentions and one name mention.

```
<?xml version="1.0" encoding="UTF-8"?>
<gann>
<entity id="[entity-id]" type="/Any/[entity-type]/[entity-subtype]" class="[entity-class]"
        comment="[comment]" annotator="[annotator-id]">
        <mention id="[mention-id]" span="[extent-range]" head="[head-ranges]" type="[mention-type]"
        role="[mention-role]" metonymic="[metonymic-type]" comment="[comment]"
        annotator="[annotator-id]"/> <!-- text="[text]" head="[head]" -->
        <mention [...]>
        [...]
        </entity>
        <entity [...]>
        [...]
        </entity>
        [...]
        </entity
        </entity>
        [...]
        </entity>
```

Figure 3: General format of a gann annotation file.

```
<entity id="10"
  type="/Any/Organization/Commercial"
  class="SPC" comment="" annotator="1">
  <mention id="22" span="497..510"
     head="506..510" type="NOM" role="N/A"
     metonymic="false" comment="" annotator="1"
     /> <!-- text="a Soviet bank" head="bank" -->
    <mention id="23" span="552..589"
     head="552..589" type="NAM" role="N/A"
     metonymic="false" comment="" annotator="1"
     /> <!-- text="the Bank for Foreign Economic
     Affairs" head="Bank for Foreign Economic
     Affairs" -->
  <mention id="24" span="642..655"
     head="651..655" type="NOM" role="N/A"
     metonymic="false" comment="" annotator="1"
     /> <!-- text="a Soviet bank" head="bank" -->
</entity>
```

### 4. Discussion on the Usefulness of C-3

As pointed out earlier, discourse coherence and coreference are interrelated phenomena (Hobbs, 1979) as discourse relations are defined based on entities, their properties and the events in which they participate (Wolf & al., 2003). A resource that has both phenomena annotated on the same text can be used to test hypotheses about their interrelation, or to detect other phenomena. Even while coreference is considered the "easier" task between the two, the availability of human-annotated coreference information eliminates the errors introduced by an automated coreference detector and thus clarifies the contribution of a larger system that uses coreference as a tool for a different task. The task for which C-3 was designed and to which it was applied is discourse relation detection, in particular Discourse GraphBank elaboration detection.

In the system implemented for elaboration detection, the new coreference knowledge was most of all useful as an indicator of the semantic classes of the detected discourse relations. Annotated entity types and subtypes and mention roles loosely correspond to elaboration classes. For example, an elaboration done by a discourse segment on an entity of type ORG/Commercial and an entity of type GPE/State-or-Province with a GPE.LOC mention role is likely to be an "elab-org-loc", an elaboration on an organization and a location. The mapping between annotated elements and relation classes was not one-to-one and was quite imperfect, but the clues offered by the annotated properties were the most valuable in detecting the elaboration class.

Secondly, mention types (e.g., PRO, NAM or APP) were important to signal relations. APP and ARC constructions are nearly always elaborations, and pronouns are strong indicators of elaborations in certain context patterns.

A third way in which the coreference annotation proved useful for its intended task was in pointing out potential connections between segments, especially in the case of long-distance relations. If far away discourse segments manifested coreference between some of their entities, their connection was worth exploring.

Finally, the position of the mentions in their coreference chains influenced the choice of the nuclei and satellites of the proposed relations and the direction in which they connected: the first appearance of a coreferring mention in a segment can signal it as a nucleus, some relations prefer to connect back to the first appearance of a mention, and nuclei usually appear before satellites in the discourse.

In the other direction, the coherence annotations can be of use when trying to detect the coreference relations, in the vein of (Hobbs, 1979). The newly annotated coreference relations can serve as a standard to compute the accuracy of the detected coreference. When two discourse segments are known to be in a relation, coreference is expected to occur between their entities in order for them to fit the definition of the discourse relation in question. In many of his works, Hobbs declared coreference a byproduct of coherence, and the availability of coherence information can easily serve to verify this theory.

Due to the separation between annotation and text, the C-3 annotation task did not take into account the existing coherence annotation, and the necessity of consulting the

existing annotation was minimal. However, the two sets of annotations can be used to improve each other. To improve the coherence annotation knowing the coreference information, entity types, subtypes and mention roles can be compared against annotated discourse relation classes in order to bring out inconsistencies or mislabeling. For example, an elaboration might be labeled as "elab-per" while the entity and mention annotation shows no person in the satellite; this elaboration annotation would then be slated for review. For the reverse, two segments that are annotated as being in a discourse relation could highlight overlooked coreference relations between their entities, or two segments that are not marked as being in a relation could make an annotator reassess the coreference between their entities. By taking into account the strengths and weaknesses of both annotations and their interconnection, a reliable corpus version can be consolidated as future work.

### 5. Conclusions

We presented a new coreference and coherence corpus (C-3) obtained by annotating Discourse GraphBank with coreference information, and a new coreference annotation tool with a simple to use interface. We described the base corpus and the annotation guidelines used, detailing the differences between them and the ACE guidelines. We introduced the annotation tool, the annotation procedure and the format of the annotation files. We ended with a discussion of the potential uses and applications for the new corpus. The knowledge value this resource brought to a relation detection system suggests it is a valuable addition to the plethora of existing natural language resources.

## 6. Acknowledgements

We wish to thank Daniel Worlton for his assistance and valuable contribution in the early stages of this annotation project, and to the three anonymous reviewers for their helpful and constructive criticism of the abstract version of this paper.

### 7. References

- ACE. (2004). Annotation Guidelines for Entity Detection and Tracking (EDT) Version 4.2.6 200400401. Linguistic Data Consortium, University of Pennsylvania.
- Carlson, L., Marcu, D., Okurowski, M. (2002). *RST Discourse Treebank*. Philadelphia, PA: Linguistic Data Consortium.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37-46.
- DeCristofaro, J., Strube, M., McCoy, K.E. (1999). Building a tool for annotating reference in discourse. In ACL '99 Workshop on the Relationship between Discourse~Dialogue Structure and Reference, University of Maryland, Maryland, 21 June, 1999, pp. 54-62.

- Hobbs, J. (1979). Coherence and Coreference. *Cognitive Science*, Vol. 3, No. 1, pp. 67-90.
- Hobbs, J. (1985). On the coherence and structure of discourse. Stanford, CA.
- Kehler, A. (2002). Coherence, reference, and the theory of grammar. Stanford, CA.
- Mann, W.C., Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization.
- MUC-6. (1995). Coreference task definition.
- Müller, C., Strube, M. (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In: Sabine Braun, Kurt Kohn, Joybrato Mukherjee (Eds.): *Corpus Technology* and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214. (English Corpus Linguistics, Vol.3)
- Prasad, R, Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. *LREC*, 2008.
- Wolf, F., Gibson, E., Fisher, A., Knight, M. (2003). A procedure for collecting a database of texts annotated with coherence relations. Technical report, Massachusetts Institute of Technology.