A procedure for collecting a database of texts annotated with coherence relations

Florian Wolf, Edward Gibson, Amy Fisher, Meredith Knight

December 1, 2003
Table of Contents

A procedure for collecting a database of texts annotated with coherence relations ........................................ 1
1 Introduction.................................................................................................................................................. 3
2 Comparison with the RST Treebank......................................................................................................... 4
3 Data structures ........................................................................................................................................ 6
   3.1 Basic assumptions.......................................................................................................................... 6
   3.2 Defining discourse segments ....................................................................................................... 7
4 Coherence relations and annotation procedure .................................................................................... 9
   4.1 Overview of coherence relations.................................................................................................. 9
   4.2 Definitions of coherence relations ............................................................................................ 10
      4.2.1 Resemblance relations ....................................................................................................... 10
      4.2.2 Cause-Effect relations ...................................................................................................... 13
      4.2.3 Temporal Sequence relation ............................................................................................. 14
      4.2.4 Attribution relation ............................................................................................................ 15
      4.2.5 Same relation .................................................................................................................... 15
5 Annotation tools and file formats .......................................................................................................... 15
   5.1 File formats .................................................................................................................................. 15
   5.2 Java annotator tool ..................................................................................................................... 16
   5.3 Perl scripts.................................................................................................................................... 19
      5.3.1 Perl script annotator2hierarchical.pl ................................................................................ 19
      5.3.2 Perl script hierarchical2annotator.pl ............................................................................... 20
   5.4 File name standards....................................................................................................................... 21
6 References.............................................................................................................................................. 22
7 Appendix A –Annotation procedure “recipes”....................................................................................... 23
   7.1 Connectives that help in determining coherence relations .......................................................... 23
   7.2 Important distinctions .................................................................................................................. 23
   7.3 General points ............................................................................................................................ 23
1 Introduction

Consider the following two passages from Jurafsky & Martin (2000):

(1) coherent
   (1a) Bill hid John’s car keys.
   (1b) He was drunk.

(2) incoherent
   (2a) Bill hid John’s car keys.
   (2b) He likes spinach.

Whereas Example (1) is a coherent sequence of sentences, Example (2) is not. The sentences in Example (1) can be related to each other in the following way: John was drunk, which is why Bill did not want him to drive and therefore Bill hid John’s car keys. By contrast, establishing any such relation between the two sentences in Example (2) is much harder. This is why Example (1) is coherent, whereas Example (2) is not.

The relation between the sentences in Example (1) is causal. In addition to causal relations, there are other ways in which sentences can relate to each other (coherence relations), in their basic definitions dating from Aristotle (cf. Hobbs et al. (1993)). Other coherence relations include similarity or contrast relations, like between sentences (3a) and (3b) in Example (3). Sentences might also elaborate on other sentences, as in Example (4), where sentences (4b) and (4c) both elaborate on sentence (4a) (notice also that sentences (4b) and (4c) are in a similarity / contrast relation):

(3) Contrast relation
   (3a) John likes ice cream.
   (3b) Matt prefers cheesecake.

(4) Elaboration relations
   (4a) Fruit are some of John’s favorite kind of food.
   (4b) He especially likes apples.
   (4c) However, he also likes kiwis a lot.

Systematic analyses of these phenomena are crucial to the investigation of human communication; virtually any form of human communication involves multiple clauses that are in some relation to each other. Furthermore, coherence relations can affect aspects of human language processing, such as pronoun resolution (Hobbs (1979); Kehler (2002); Wolf et al. (2003)). In addition, a better understanding of text coherence could improve any natural language engineering application that requires access to informational structures of texts. Examples are information retrieval, text summarization, and machine translation.

In order to allow systematic analyses of text coherence, a database of texts annotated with coherence relations has been collected. All types of coherence relations used in the annotations will be defined in detail in Section 4.2. A plan for the future is to also annotate information about
anaphoric relations, words that explicitly signal coherence relations (“because”, “although”, etc),
and inter-sentential lexical relations.

The database will be designed such that the different kinds of information can be stored in
separate but linked files (one file for coherence relations, one for anaphoric relations, etc). Such a
modular design will facilitate later addition of more information to the database, for example parts
of speech or (partial) syntactic structures. Such additional information can then be represented in
additional files, making it unnecessary to edit already existing files. Furthermore, the tools used for
analysis of the data can then be modified easier as well. Details about the structure of the files as
well as about how the files are linked will be given in Section 5.4.

The text material used in the present project is raw unparsed text from the AP Newswire,
the Wall Street Journal, and GRE and SAT texts. The texts deal with a wide range of topics
(politics, finance, sports, entertainment, etc). Table 1 shows corpus statistics for words and
discourse segments (cf. Section 3.2) for 135 annotated texts.

<table>
<thead>
<tr>
<th>Number of words</th>
<th>Number of discourse segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>mean 545</td>
</tr>
<tr>
<td>min</td>
<td>min 161</td>
</tr>
<tr>
<td>max</td>
<td>max 1409</td>
</tr>
<tr>
<td>median</td>
<td>median 529</td>
</tr>
</tbody>
</table>

Table 1. Corpus statistics.

Each text was independently annotated by two annotators. In order to determine inter-
annotator agreement, we constructed a confusion matrix of coherence relations in both annotations
(the columns of the confusion matrix are the coherence relations assigned by one annotator; the
rows are the coherence relations assigned by the other annotator). For all annotations of the 135
texts, the agreement is 88.45%, per chance agreement is 24.86%, and kappa is 84.63%. There were
no systematic disagreements between annotators, and no systematic differences depending on text
length or number of arcs in an annotation graph: annotator agreement did not differ as a function of
text length, number of arcs in a coherence graph, arc length, or kind of coherence relation.

2 Comparison with the RST Treebank

The only other existing database of texts annotated with coherence relations is the RST Discourse
Treebank (Carlson et al. (2002)). Carlson et al used an annotation scheme that was based on
Rhetorical Structure Theory (RST; Mann & Thompson (1988)). However, a problem with the RST
Discourse Treebank is that it assumes tree graphs to represent coherence relations. It can be shown
that this assumption does not hold, since graphs representing coherence structures contain crossed
dependencies (Wolf & Gibson (2003)). Consider the following examples:
(5) **Crossed dependency I**

(5a) There is a Eurocity train on Platform 1.
(5b) Its destination is Rome.
(5c) There is another Eurocity on Platform 2.
(5d) Its destination is Zürich.

The following coherence relations hold between the sentences of this text (cf. Section 4.2 for definitions of the coherence relations):

(5b) -> (5a) elaboration
(5a) <-> (5c) parallel
(5d) -> (5c) elaboration
(5b) <-> (5d) contrast

As a figure\(^1\) (the colors of the edges are used in the Java annotation tool to represent different coherence relations, cf. Section 5.2):

![Coherence graph for Example (5).](image)

Here is another example of crossed dependencies:

(6) **Crossed dependency II**

(6a) The first planet we saw through the telescope was Jupiter.
(6b) After that, we saw Saturn.
(6c) Then we took a look at Neptune
(6d) and towards the end of the night we even saw Uranus.
(6e) In everyone’s opinion, Jupiter was the most exciting with its cloud bands and the moons.
(6f) Saturn’s ring was fun to see, too,
(6g) but both Neptune and Uranus seemed just like two little white dots.

Figure 2 represents the coherence relations between the sentences in Example (6).

---

\(^1\) To improve “legibility” of the figure, the undirected edges **Parallel** and **Contrast** are represented as such in this figure, and not as cycles of directed edges.
Figure 2. Coherence graph for Example (6).

The boxes in the figure represent a coherence relation applying to groups of sentences. In Example (6), for example, sentence (6g) elaborates on both sentences (6c) and (6d). Furthermore, sentence (6g) is in a contrast relation with sentences (6e) and (6f).

The crossed dependencies in both examples cannot be represented in a tree. Preliminary results indicate that such crossed dependencies are in fact abundant in texts. Section 3.1 explains the data structure used in the present project in more detail.

3 Data structures

3.1 Basic assumptions

The following assumptions are made about the data structure that represents coherence relations in texts:

- The data structure is a directed graph where nodes represent discourse segments and groups of discourse segments (henceforth DSs), and labeled directed arcs represent coherence relations holding between the DSs and groups of DSs.

- DSs are non-overlapping units of text (cf. Section 3.2 for more detailed definitions).

- Groups of DSs are connected subgraphs of a coherence graph.

- A graph representing a coherent text is connected. An unconnected graph implies that the underlying text is not fully coherent and that it contains discourse segments that do not relate to any other discourse segment in the text.

- There are symmetrical and asymmetrical coherence relations (cf. Marcu (2000)):
  - In symmetrical coherence relations, the DSs involved in the coherence relation play equally important roles in the text. For example, similarity / contrast relations are symmetrical relations. Symmetrical relations are represented as cycles of identical, labeled, directed arcs.
In asymmetrical coherence relations, one DS plays a more important role in the text than the other. For instance, in elaboration relations, the elaborating DS plays a less important role in the text than the general DS that is elaborated. In asymmetrical relations, the arcs go from the less important DS (the Satellite) to the more important DS (the Nucleus).

- Except cycles representing symmetrical coherence relations, any two nodes are related by a unique coherence relation / labeled edge.
- One node can relate to more than one other node.
- Groups of DSs should only be assumed if otherwise truth conditions are changed. The following passage is an example where truth conditions are changed if no groups of DSs are assumed:

  1. Arizona usually has very pleasant weather.
  2. Only sometimes it gets unpleasant
  3. but only if there are clouds.

In this example, the truth condition of DS 2 alone is different from the truth condition of DSs 2 and 3 together. DS 2 alone would allow one to say that the weather is unpleasant if it is hot and there are no clouds. However, DSs 2 and 3 together contradict that assertion. By contrast, the following example does not require groups of DSs to preserve truth conditions:

  1. Five stocks went down last Friday.
  2. For example, Cisco’s stock lost ten percent.
  3. The Cisco CEO voiced his concern about this development.

Here, it is enough to relate only DS 2 to DS 1. DSs 2 and 3 are related, but DS 3 does not necessarily participate in the relation of DSs 1 and 2. Therefore no group of DSs including DSs 2 and 3 should be assumed here.

- If a DS \(d_0\) modifies a DS \(d_1\) which modifies a DS \(d_2\) or group of DSs \(d_{2-n}\), no inheritance is assumed from \(d_0\) to \(d_2\) or \(d_{2-n}\).

- If a DS \(d_0\) is modified by a (group of) DSs \(d_{1-k}\) (with \(k \geq 1\)) and if \(d_0\) modifies a DS \(d_m\) (\(m > k\)) or a group of DSs \(d_{m-n}\) (\(n > m > k\)), no inheritance is assumed from \(d_{1-k}\) to \(d_m\) or \(d_{m-n}\).

- If a DS \(d_0\) and a DS \(d_1\) are in a Resemblance or Contrast relation and if \(d_0\) and \(d_1\) both modify a DS \(d_2\) or a group of DSs \(d_{2-n}\), there have to be arcs both from \(d_0\) and \(d_1\) to \(d_2\) or \(d_{2-n}\).

### 3.2 Defining discourse segments

Most researchers agree that discourse segments are non-overlapping units of text (cf. Marcu (2000); Polanyi (1996); but see Wiebe (1994)). However, it is much less clear how exactly such non-
overlapping discourse segments are defined or delimited. Examples (1)-(4) from Marcu (2000) show that there is not necessarily always a one-to-one match between syntactic and discourse segments. While (1)-(4) all express basically the same discourse segments (connected by a Cause-Effect relation) the syntactic boundaries differ, especially between (1)-(3) and (4).

(7) [ Xerox Corp.’s third-quarter net income grew 6.2% on 7.3% higher revenue. ] [ This earned mixed reviews from Wall Street analysts. ]
(8) [ Xerox Corp.’s third-quarter net income grew 6.2% on 7.3% higher revenue, ] [ which earned mixed reviews from Wall Street analysts. ]
(9) [ Xerox Corp.’s third-quarter net income grew 6.2% on 7.3% higher revenue, ] [ earning mixed reviews from Wall Street analysts. ]
(10) [ The 6.2% growth of Xerox Corp.’s third-quarter net income on 7.3% higher revenue earned mixed reviews from Wall Street analysts. ]

As a basic rule, discourse segments (DSs) here will be assumed to be

- clauses delimited by commas or full-stops, since commas and full-stops are assumed to be equivalents of phrase boundaries in speech (cf. Hirschberg & Grosz (1992))
- elements of text (especially modifiers) that are separated by commas. The idea here is that commas that are equivalent to intonational phrase boundaries in speech should denote DSs.
- attributions, as in “John said that…”. This is empirically motivated. The texts used here are taken from news corpora, and there, attributions can be important carriers of coherence structures. For instance, consider a case where some Source A and some Source B both comment on some Event X. It should be possible to distinguish between a situation where Source A and Source B make basically the same statement about Event X, and a situation where Source A and Source B make contrasting comments about Event X.

Here are some refinements of these basic rules:

- Clauses delimited by commas or full-stops are DSs. Commas are not DS-boundaries if they separate elements of a complex NP, or in cases like the following:
  - [ It wasn’t known to what extent, if any, the facility was damaged. ] (Marcu (2000))
- Elaborations (cf. Section 3.1.1 on MUC-7 annotation tags) are separate DSs:
  - [ Mr. Jones, ][ spokesman for IBM, ] [ said... ]
- Infinitival clauses are separate DSs (to has to be substitutable by in order to):
  - [ The arm can be fitted to allow it to grasp, lift and turn objects of differing sizes ]
  - [ to suit a variety of tasks. ]
• Infinitival complements of verbs are not treated as separate DSs:
  \[ \text{The machinery is of the type used to make small parts in metal cutting shops.} \]  
  (Marcu (2000))

• Participial complements of verbs are not treated as separate DSs:
  \[ \text{The company misled many customers into purchasing more credit-data services.} \]  
  (Marcu (2000))

• Gerund forms that are clausal modifiers are treated as DSs:
  \[ \text{the prices benefited from price reductions arising from introduction of the consumption tax} \]

• Prepositional phrases are treated as DSs if they are clausal modifiers:
  \[ \text{With the ground stone being laid, they were able to move on.} \]

• Whenever a source for a statement is mentioned, the statement and the source are treated as separate DSs.
  \[ \text{"Gorbachev deserves more credit than Reagan does," Thomas Cronin said.} \]

• DSs can contain ellipses (elided part in bold):
  \[ \text{Human workers remain responsible for keeping inventory and coordinating different aspects of the production line.} \]

• Time-, space-, personal- or detail-elaborations are treated as DSs:
  \[ \text{This past year, the original robot was replaced with one able to perform more tasks.} \]
  \[ \text{Andy Russell, a spokesman for IBM} \]

• Strong discourse markers (e.g. because, although, after, while) are assumed to delimit DSs:
  \[ \text{IBM will benefit because we will be helping to train the (computer-integrated manufacturing) workers and decision makers of today and tomorrow.} \]

4 Coherence relations and annotation procedure

4.1 Overview of coherence relations

The coherence relations used are from Hobbs (1985) and Kehler (2002), with a few exceptions (noted). The coherence relations are illustrated in Figure 3 (the colors are used in the Java annotation tool to represent different coherence relations, cf. Section 5.2). Notice, however, that the hierarchy of coherence relations implied in Figure 3 only serves illustrative purposes and does not imply a certain type hierarchy of coherence relations.
4.2 Definitions of coherence relations

4.2.1 Resemblance relations

Resemblance relations establish commonalities and contrasts between corresponding (sets of) discourse entities or properties (Kehler (2002)). Corresponding (sets of) discourse entities or properties are usually syntactically and / or semantically parallel. Parallel and Contrast relations are symmetrical. Here this is represented by a cycle of directed edges. By contrast, Exemplification, Generalization and Elaboration relations are asymmetrical. They have a Satellite and a Nucleus.

4.2.1.1 Parallel

Tag: par
Relation type: symmetrical
Definition: Infer a set of entities from DS₀, E(DS₀), and a set of entities from DS₁, E(DS₁). Then infer commonalities between members of E(DS₀) and E(DS₁).

Example:

*John organized rallies for Clinton, and Fred distributed pamphlets for him.*

→ “organize” and “distribute” correspond (although they are not in a synonym relation², and have a common superclass (e.g. “support a political candidate”). The arguments of these predicates – “Clinton” and “him” respectively - also correspond

---

² The relevant synonym and antonym relations will be taken from WordNet 1.6
Parallel-B

Tag: parallel
Relation type: symmetrical
Definition: Two groups of DSs are in a parallel relation (the “Parallel” relation described in Section 4.2.1.1 is a parallel relation between two single DSs).

Example:
[ The university spent $30,000 to upgrade lab equipment in 1987. An estimated $60,000 to $70,000 was earmarked in 1988. ]
[ International Business Machines Corp. recently pledged $1.2 million in computer equipment and software to the university as part of an IBM program to aid 48 college-based robotics labs across the country. ]

4.2.1.2 Contrast

Tag: contr (same tags for Contrast-1 and Contrast-2)
Relation type: symmetrical
Definition: Infer a set of entities from DS0, E(DS0), and a set of entities from DS1, E(DS1). Then infer contrasts between members of E(DS0) and E(DS1).
- Contrast-1 is a contrast between corresponding predicates in DS0 and DS1. The arguments of these contrasting predicates are identical.
- Contrast-2 is a contrast between the arguments of corresponding predicates in DS0 and DS1. The predicates over these contrasting arguments are identical.

Examples:
Contrast-1: John supported Clinton, but Mary opposed him.
→ antonym-relation between the predicates, “support” and “oppose”
Contrast-2: John supported Clinton, but Mary supported Bush.
→ contrast between the arguments of predicates – “support(Clinton)” and “support(Bush)”

Contrast-B

Tag: contrast
Relation type: symmetrical
Definition: two groups of DSs are in a contrast relation (the “Contrast” relation described in Section 4.2.1.2 is a contrast relation between two single DSs).

Example:
[ Alan Spoon, recently named Newsweek president, said Newsweek's ad rates would increase 5% in January. A full, four-color page in Newsweek will cost $100,980. ]
[ In mid-October, Time magazine lowered its guaranteed circulation rate base for 1990 while not increasing ad page rates; with a lower circulation base, Time's ad rate will be effectively 7.5% higher per subscriber; a full page in Time costs about $120,000. ]
4.2.1.3 Example

Tag: examp
Relation type: asymmetrical – example = Satellite, exemplified = Nucleus

Definitions:
- Infer a set of entities from DS₀, E(DS₀), and a set of entities from DS₁, E(DS₁). Then find some element in E(DS₁) that is a member or subset of the corresponding element in E(DS₀).
- Infer a set of entities from DS₀, E(DS₀), and a set of entities from DS₁, E(DS₁). Then find some element in E(DS₁) that is a new instantiation of an entity in E(DS₀).

Example:
Young aspiring politicians often support their party's presidential candidate. For instance, John campaigned hard for Clinton in 1992.

→ “John” is in E(DS₀) and it is a member of “young aspiring politicians”, which is the corresponding element in E(DS₀).

→ “John” is also a new instantiation of “young aspiring politicians”, which is an entity in E(DS₀).

4.2.1.4 Generalization

Tag: gen
Relation type: asymmetrical – example = Satellite, generalization = Nucleus

Definition:
- Infer a set of entities from DS₀, E(DS₀), and a set of entities from DS₁, E(DS₁). Then find some element in E(DS₀) that is a member or subset of the corresponding element in E(DS₁).
- Infer a set of entities from DS₀, E(DS₀), and a set of entities from DS₁, E(DS₁). Then find some element in E(DS₀) that is a new instantiation of an entity in E(DS₁).

Example:

→ “John” is in E(DS₀) and it is a member of “young aspiring politicians”, which is the corresponding element in E(DS₁).

→ “John” is also a new instantiation of “young aspiring politicians”, which is an entity in E(DS₀).

4.2.1.5 Elaboration

Tag: elab
Relation type: asymmetrical – elaboration = Satellite, elaborated = Nucleus

Definition: Infer a set of coherent entities, E(DS₀, DS₁) from DS₀ and DS₁. The members of E(DS₀, DS₁) are centered around a common event or entity, e₀₁.
Example:

A young aspiring politician was arrested in Texas today. John Smith, 34, was nabbed in a Houston law firm while attempting to embezzle funds for his campaign.

→ “arrested(young aspiring politician)”, “John Smith”, “Houston law firm”, “campaign funds” etc. are a set of coherent entities, centered around a common event, arrest(politician).

Subclasses of Elaboration (cf. Chinchor (1997)):

- **Organization** – org
  The Satellite gives information about an organization involved in the event described by the Nucleus

- **Person** – pers
  The Satellite gives information about a person involved in the event described by the Nucleus

- **Location** – loc
  The Satellite gives information about the location where the Nucleus took place

- **Time** – time
  The Satellite gives information about the time at which the Nucleus took place

- **Number** – num
  The Satellite gives information about the time at which the Nucleus took place

- **Detail** – det
  The Satellite gives details about an entity involved in the event described by the Nucleus. The details cannot be captured by any of the relations above.

An elaborating DS can also include more than one of these subclasses. In that case, all subclasses should be annotated (e.g. elab-time-loc).

4.2.2 Cause-Effect relations

Caused-Effect relations establish a causal inference path between discourse segments. They are directed, i.e. there is a Satellite (Cause) and a Nucleus (Effect).

4.2.2.1 Explanation (standard Cause-Effect relation)

Tag: ce

Relation type: asymmetrical – cause = Satellite, effect = Nucleus

Definition: Infer a causal relation between DS\textsubscript{0} and DS\textsubscript{1}. 

Examples:

Bill is a politician, and therefore he is dishonest.
Bill is dishonest because he's a politician.

→ being a politician is a reason for being dishonest.

4.2.2.2 Violated Expectation

Tag: expv
Relation type: asymmetrical – cause = Satellite, effect = Nucleus
Definition: Infer that normally there is a causal relation between DS$_0$ and DS$_1$ but that causal relation is absent between DS$_0$ and DS$_1$.

Examples:

Bill is a politician, but he's honest.
Bill is honest, even though he's a politician.

(being a politician is a reason for being dishonest, but here this causal relation is absent)

4.2.2.3 Condition

Tag: cond
Relation type: asymmetrical – condition = Satellite, result = Nucleus
Definition: the event described in the Nucleus can only take place if the event described in the Satellite also takes place (before or simultaneously with the event described in the Nucleus)

Example:

If the system works, everyone will be happy.

(everyone will only be happy if the system works, not otherwise).

4.2.3 Temporal Sequence relation

Tag: temp
Relation type: asymmetrical – first event = Satellite, second event = Nucleus
Definition: Infer a temporal sequence of the events described by DS$_0$ and DS$_1$. There is no causal relation between DS$_0$ and DS$_1$. If there is a causal relation, the relation between DS$_0$ and DS$_1$ should be described as a Cause-Effect relation.

Examples:

John bought a book. Then he bought groceries.

(there is a temporal sequence between the events described by DS$_0$ and DS$_1$, but no causal relation.)

John bought groceries. But before that he bought a book.

(there is a temporal sequence between the events described by DS$_0$ and DS$_1$, but no causal relation. The order of narration is the reverse order of event occurrence.)
4.2.4 Attribution relation

Tag: attr
Relation type: asymmetrical – attribution = Satellite, quote = Nucleus
Definition: The Satellite attributes the Nucleus to a source.

Examples:

John said that...
According to John,...

4.2.5 Same relation

Tag: same
Relation type: symmetrical
Definition: a DS has intervening material; the “Same” relation is no coherence relation, but a “trick” that allows dealing with DSs nested in other DSs.

Example:

The economy, according to the G-8 countries, should improve by early next year.
(the underlined material is in a “Same” relation)

5 Annotation tools and file formats

5.1 File formats

Consider again the sequence from Section 2:

1. There is a Eurocity train on Platform 1.
2. Its destination is Rome.
3. There is another Eurocity on Platform 2.
4. Its destination is Zürich.

As pointed out in Section 2, the following coherence relations hold between the DSs of this text:

2 -> 1 elaboration
1 <-> 3 parallel
4 -> 3 elaboration
2 <-> 4 contrast

Using the annotator tool (cf. Section 5.2) would produce a text file that looks like this:

2 2 1 1 elab-det
1 1 3 3 par
4 4 3 3 elab-det
2 2 4 4 contr
The first two numbers in each line mark the group of DSs that are the satellite of a coherence relation. For example, in the first line, “2 2” indicates that the satellite of the “elab-det” coherence relation starts at DS 2 and also ends at DS 2. Also in the first line, “1 1” indicates that the nucleus of the “elab-det” coherence relation starts at DS 1 and also ends at DS 1.

Notice that coherence relations with no satellite or nucleus, such as parallel or contrast, are annotated as if they had a satellite or a nucleus. However, for further processing, these coherence relations will be reverse-duplicated. For example, the parallel relation from line 2 in the text above would be represented as a cycle. This is a workaround to avoid having to deal with mixed graphs that contain both directed and undirected edges.

\[1 1 3 3 \text{par} \quad // \text{this line is in the annotation file}\]
\[3 3 1 1 \text{par} \quad // \text{this line is the reverse-duplicated relation}\]

These annotation text files could also be translated into XML format. The XML-based annotation scheme could for instance be modeled after Bird & Liberman (2000).

5.2 Java annotator tool
The Java tool annotator is used for the coherence relation annotation. Figure 4 shows a screenshot of the annotator tool. Its functions include

- discourse annotation (saves annotation files in the format described in Section 5.1)
- breadth-first graph traversal of the annotation structure to check if the coherence graph constructed thus far is connected
- detection of crossed dependencies (including an option to save the results as a file (text-number)-crossed-dependencies)
- save complete coherence graphs or parts of coherence graphs as Postscript files
- colored edges representing coherence relations:
  - green:
    - Parallel (par and parallel)
    - Contrast (contr and contrast)
  - blue:
    - Exemplification
    - Generalization
    - Elaboration (including subclasses)
  - red:
    - Cause-Effect
    - Violated Expectation
    - Condition
  - cyan:
    - Temporal Sequence
- orange:
  - Attribution
- Indicating groups of DSs, colored according to the coherence relation they participate in (cf. colors above)
Figure 4. Screenshot of the annotator tool.
5.3 Perl scripts

This section describes Perl scripts for further data processing.

5.3.1 Perl script annotator2hierarchical.pl

- **Input:** [text-number]-annotation
- **Output:** [text-number]-hierarchical-annotation

The `annotator2hierarchical.pl` script converts `annotator` output files to a format that better takes into account the hierarchical structure of the coherence graph. Furthermore, cycles are added for symmetrical coherence relations. Below is an example of a conversion.

- **Input to `annotator2hierarchical.pl`:**
  - **Output of `annotator` tool:**
    1 1 0 0 elab
    3 3 2 2 par
    2 2 0 1 elab
    4 4 5 5 temp
    4 4 0 3 elab
    5 5 1 1 elab
  - **Graphical representation in `annotator` tool:**
    ![Graphical representation](image-url)
• **Output of annotator2hierarchical.pl:**
  o **Text format:**
    
    ```
    1 0 elab
    3 2 par
    2 3 par
    2 group-0-1 elab
    4 5 temp
    4 group-0-3 elab
    0 group-0-1 group
    1 group-0-1 group
    group-0-1 group-0-3 group
    2 group-0-3 group
    3 group-0-3 group
    ```
  o **Graphical representation:**

  ![Graphical representation](image)

  This hierarchical representation of the coherence graph facilitates hierarchical pattern searches and is a better representation of nested groups. Notice that the goal is not to convert the output of the annotator tool into a tree structure – crossed dependencies are maintained in the hierarchical representation created by annotator2hierarchical.pl.

5.3.2 **Perl script hierarchical2annotator.pl**

- **Input:** [text-number]-hierarchical-annotation
- **Output:** [text-number]-annotation

This Perl script does the reverse of annotator2hierarchical.pl (it converts hierarchical annotations into annotator format).
5.4 File name standards

- \[text-number\] – raw text file, text segmented into DSs
- \[text-number\]-annotation – annotation for a text file, created with the annotator tool
- \[text-number\]-hierarchical-annotation – annotation file created with annotator2hierarchical.pl
6 References
7 Appendix A – Annotation procedure “recipes”

7.1 Connectives that help in determining coherence relations
Following suggestions by Hobbs (1985) and Kehler (2002), in order to help with determining coherence relations, try to connect the DSs under consideration with one of the words in the table below:

<table>
<thead>
<tr>
<th>Coherence Relation</th>
<th>Connective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cause-Effect</td>
<td>Because</td>
</tr>
<tr>
<td>Violated Expectation</td>
<td>Although</td>
</tr>
<tr>
<td>Condition</td>
<td>if…then</td>
</tr>
<tr>
<td>Parallel</td>
<td>(and) similarly</td>
</tr>
<tr>
<td>Contrast</td>
<td>by contrast</td>
</tr>
<tr>
<td>Temporal Sequence</td>
<td>and then</td>
</tr>
<tr>
<td>Attribution</td>
<td>according to…</td>
</tr>
<tr>
<td>Example</td>
<td>for example</td>
</tr>
<tr>
<td>Elaboration</td>
<td>also, furthermore, in addition</td>
</tr>
<tr>
<td>Generalization</td>
<td>in general</td>
</tr>
</tbody>
</table>

7.2 Important distinctions

- **Difference Example – Elaboration:** an Example sets up an additional entity (the example), whereas an Elaboration gives more detail about an already existing entity (the one on which one elaborates)

- **Difference Nucleus – Satellite:** If one had to summarize the text: the Nucleus is what would have to remain in the text in order for the text to still be comprehensible, the Satellite is what could be left out.

7.3 General points

- **Inferences:** In doubt, use a coherence relation that requires less inferences (inferences are basically assumptions one makes about things or facts that are not explicitly given in the text)

- **Long-distance dependencies:** When connecting non-adjacent DSs, make sure that they really go together. Imagine them being immediately adjacent. That should create a coherent sequence of sentences.