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

Florian Wolf

October 3, 2003

Send correspondence to:
Florian Wolf
Massachusetts Institute of Technology
Department of Brain and Cognitive Sciences
MIT NE20-448
3 Cambridge Center
Cambridge, MA 02139
Ph.: ++1 617 452 2474
Email: fwolf@mit.edu
http://web.mit.edu/~fwolf/www

Table of Contents

1 Introduction

Consider the following two passages from Jurafsky & Martin (2000: 695):

(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, 1990). 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 et al, 1990; Kehler, 2002; Wolf, Gibson & Desmet, 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. In addition to annotating the coherence relations that hold

between the sentences in the texts, information relevant for determining coherence relations has been annotated for some of the texts. This includes information about anaphoric relations as well as information about words or phrases that explicitly signal coherence relations (so-called *connectives*, words like "because", "although", etc). A plan for the future is to also annotate information about inter-sentential lexical relations.

The database will be designed such that the different kinds of information are 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 7.5.

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.

number of words		**number of discourse segments**	
mean	545	mean	61
min	161	min	6
max	1409	max	143
median	529	median	60

Table 1. Corpus statistics.

2 Comparison with other projects

The only other existing database of texts annotated with coherence relations is the *RST Discourse Treebank* (Carlson, Marcu & Okurowski, 2002). Carlson et al used an annotation scheme that was based on **R**hetorical **S**tructure **T**heory (RST; Mann & Thompson, 1988). However, there are two major problems with the *RST Discourse Treebank*:

- no information on anaphoric relations or other coherence-related linguistic information
- tree graphs are used to represent the coherence relations in a text

The lack of anaphoric or other coherence-related information is an issue since it is not obvious how the existing database might be extended easily to contain that kind of information. Thus, the *RST Discourse Treebank* does not, at least in its present format, allow a systematic investigation of the role that different kinds of linguistic information play in determining coherence relations. However, such a systematic investigation would be important, since various researchers have argued for the importance of anaphoric relations in determining coherence relations or vice versa (Cristea, Ide & Romary, 1998; Schauer, 2000), for the importance of explicit cues for coherence

relations (Marcu, 2000), or for the importance of lexical relations (Barzilay, 1997). One goal of the present project is to allow systematic investigations of these factors.

Another important 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 7.3):

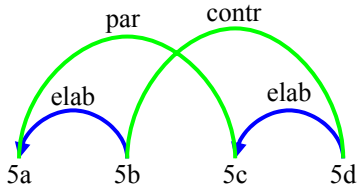


Figure 1. Coherence graph for Example (5).

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.

[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.

The following figure represents the coherence relations between the sentences in Example (6):

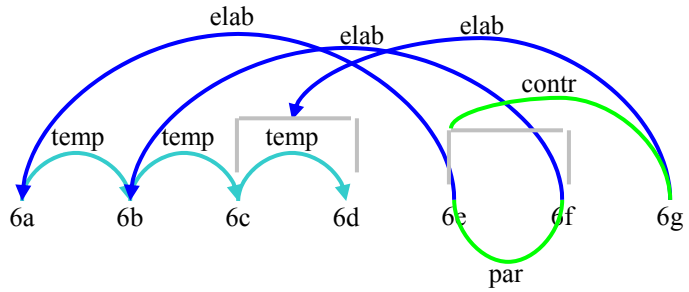


Figure 2. Coherence graph for Example (6).

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 $d_{2\text{-}n}$, no inheritance is assumed from d_0 to d_2 or $d_{2\text{-}n}$.

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

- 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 $d_{2\text{-}n}$, there have to be arcs both from d_0 and d_1 to d_2 or $d_{2\text{-}n}$.

3.2 Defining discourse segments

Most researchers agree that discourse segments are non-overlapping units of text (cf. Marcu, 2000; Polanyi & van den Berg, 1997; 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:
  - *[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:
  - *[The company misled many customers **into** purchasing more credit-data services.]* (Marcu, 2000)

- Gerund forms that are clausal modifiers are treated as DSs:
  - [*the prices benefited from price reductions*][*arising from introduction of the consumption tax*]

- Prepositional phrases are treated as DSs if they are clausal modifiers:
  - *[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.
  - *["Gorbachev deserves more credit than Reagan does,"] [Thomas Cronin said.]*

- DSs can contain ellipses (elided part in bold):
  - *[**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:
  - *[This past year,][the original robot was replaced with one able to perform more tasks.]*
  - *[Andy Russell,][a spokesman for IBM]*

- Strong discourse markers (e.g. *because*, *although*, *after*, *while*) are assumed to delimit DSs:
  - *[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 Type hierarchy of coherence relations

The coherence relations used are from Hobbs (1985) and Kehler (2002), with a few exceptions (noted). These coherence relations are in a partial hierarchical order[2]. This type hierarchy is illustrated in the figure below (the colors are used in the Java annotation tool to represent different coherence relations, cf. Section 7.3):

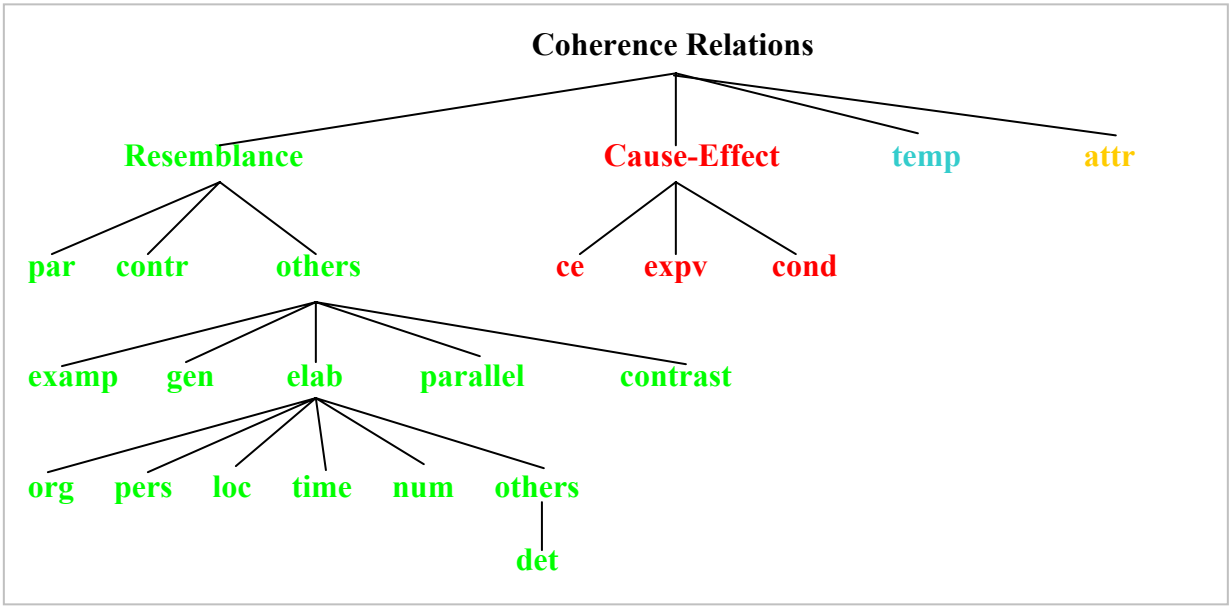


Figure 3. Type hierarchy for coherence relations.

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, 1995). 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.

[2] Notice that Hovy & Maier (1998) show that the type hierarchy of coherence relations may be open-ended vertically but is bounded horizontally. This implies that if new, more fine-grained coherence relations are introduced into the taxonomy, they will form subsets of already existing coherence relations. Therefore introducing new coherence relations (a vertical extension of the type hierarchy) should not present a problem for the taxonomy as a whole.

4.2.1.1 Parallel

Tag: par
Relation type: symmetrical
Definition: Infer a set of entities from DS_0, $E(DS_0)$, and a set of entities from DS_1, $E(DS_1)$. Then infer commonalities between members of $E(DS_0)$ and $E(DS_1)$.

Example:

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

→ "organize" and "distribute" correspond (although they are not in a synonym relation[3], and have a common superclass (e.g. "support a political candidate"). The arguments of these predicates – "Clinton" and "him" respectively - also correspond

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 DS_0, $E(DS_0)$, and a set of entities from DS_1, $E(DS_1)$. Then infer contrasts between members of $E(DS_0)$ and $E(DS_1)$.

- *Contrast-1* is a contrast between corresponding predicates in DS_0 and DS_1. The arguments of these contrasting predicates are identical.
- *Contrast-2* is a contrast between the arguments of corresponding predicates in DS_0 and DS_1. 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)"

[3] The relevant synonym and antonym relations will be taken from WordNet 1.6

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 DS_0, $E(DS_0)$, and a set of entities from DS_1, $E(DS_1)$. Then find some element in $E(DS_1)$ that is a member or subset of the corresponding element in $E(DS_0)$.
- Infer a set of entities from DS_0, $E(DS_0)$, and a set of entities from DS_1, $E(DS_1)$. Then find some element in $E(DS_1)$ that is a new instantiation of an entity in $E(DS_0)$.

Example:

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

→ "John" is in $E(DS_1)$ and it is a member of "young aspiring politicians", which is the corresponding element in $E(DS_0)$.
→ "John" is also a new instantiation of "young aspiring politicians", which is an entity in $E(DS_0)$.

4.2.1.4 Generalization

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

- Infer a set of entities from DS_0, $E(DS_0)$, and a set of entities from DS_1, $E(DS_1)$. Then find some element in $E(DS_0)$ that is a member or subset of the corresponding element in $E(DS_1)$.
- Infer a set of entities from DS_0, $E(DS_0)$, and a set of entities from DS_1, $E(DS_1)$. Then find some element in $E(DS_0)$ that is a new instantiation of an entity in $E(DS_1)$.

Example:

John campaigned hard for Clinton in 1992. Young aspiring politicians often support their party's presidential candidate.

→ "John" is in E(DS_0) and it is a member of "young aspiring politicians", which is the corresponding element in E(DS_1).
→ "John" is also a new instantiation of "young aspiring politicians", which is an entity in E(DS_0).

4.2.1.5 Elaboration

Tag: elab
Relation type: asymmetrical – elaboration = Satellite, elaborated = Nucleus
Definition: Infer a set of coherent entities, E(DS_0, DS_1) from DS_0 and DS_1. The members of E(DS_0, DS_1) are centered around a common event or entity, e_{01}.

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. MUC-7 standard, Named Entity Task Definition):

- **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

Cause-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_0 and DS_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 DS0 and DS1. There is no causal relation between DS0 and DS1. If there is a causal relation, the relation between DS0 and DS1 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 DS0 and DS1, 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 Anaphora

NP-anaphora have to be annotated manually. If a file is saved with the extension '.html', Emacs highlights the tags. NP-anaphora are pronouns as well as full NPs with an explicit antecedent (the antecedent can either be an NP or a VP).

5.1 What to annotate

The anaphor annotation guidelines here generally follow the MUC-7 coreference task definition. Exceptions will be marked. For instance, unlike in the MUC-7 coreference task, only NPs (nouns, noun phrases and pronouns) that appear more than once in a text should be annotated. That is, only anaphoric NPs should be annotated. In the following examples, entities that should be marked coreferential are printed in bold.

5.1.1 Names and other named entities

The following are examples of coreferential names and named entities:

(11) **Reuters Holding PLC** … **Reuters**
(12) Equitable of **Iowa** Cos. ... located in **Iowa**

Example (11) follows MUC-7 guidelines. But in Example (12), unlike in MUC-7, the two instances of Iowa are marked as coreferential. These instances were not marked as coreferential in MUC-7, because "Iowa" in "Equitable of Iowa Cos." is a substring of a named entity. However, MUC-7 treated named entities as atomic, so references to substrings of named entities were not possible. These considerations are not relevant to the current purpose. Therefore, references to substrings of complex NPs should be annotated.

5.1.2 Gerunds

(13) program trading, excessive spending
(14) the slowing of the economy
(15) **Slowing the economy** is supported by some Fed officials; **it** is repudiated by others.

Just as in MUC-7, references to NPs like in Examples (13) and (14) should be annotated. MUC-7 did not mark "slowing the economy" and "it" in Example (15) as coreferential. However, they should be annotated here, since the anaphoric relation supports a *Contrast* relation between the two clauses.

5.1.3 Pronouns

The annotation guidelines for pronouns follow the MUC-7 standard. First, second, and third-person pronouns are all markable. Notice that for possessives, only the possessive itself, not its arguments, are markable.

(16) **Its** chairperson
(17) "There is no business reason for **my** departure," **he** added.
(18) **He** shot **himself** with **his** revolver.

In Example (16) there are two markables, "its" and the whole NP, "its chairperson". In Example (17), "my" and "he" should be marked as coreferential. In Example (18), all three pronouns should be marked as coreferential.

5.1.4 Bare nouns

(19) The price of **aluminum** siding has steadily increased, as the market for **aluminum** reacts to the strike in Chile.
(20) **Linguists** are a strange bunch. **Some linguists** even like spinach.

As in MUC-7, instances of bare nouns such as in Example (19) should be marked as coreferential. Unlike in MUC-7, however, instances of bare nouns such as in Example (20) should also be marked as coreferential. There, the noun as well as the quantifier should be marked.

5.1.5 Implicit pronouns or gaps

(21) Bill called John and __ spoke with him for an hour.

As in MUC-7, coreference between a noun and a gap should not be marked. Notice that similarly, ellipses should also not be marked.

5.1.6 Conjoined noun phrases

(22) The **boys and girls** enjoyed **their** breakfast.

In MUC-7 the coreference between "boys and girls" and "their" is not marked (however, coreference between individual conjuncts in the conjoined NP and other phrases is marked in MUC-7). Unlike in MUC-7, here, "boys and girls" and "their" should be marked as coreferential.

5.1.7 Subsets

(23) When crack arrived in **other cities**, like **New York**, murder rates went up **there** too.

If there is a superset-subset relation between two NPs, this relation should be marked as coreferential. This justifies a coreferential relation between "other cities" and "New York". Notice that there is also a coreferential relation between "other cities" and "there".

5.1.8 Reference to clauses

(24) In 1988, **371 persons had been killed** in the nation's capital as of Dec. 30, far surpassing **the previous high total of 287**, set in 1969.

In MUC-7, nominal references to clauses are not considered markable. However, since these coreferential relations may be an important kind of anaphoric relation, they should be marked coreferential here.

5.2 How to annotate anaphora

The annotation format used here is very similar to the one used in MUC-7.

5.2.1 Annotation tags

- **n** – nouns
- **npro** – pronouns
- **min** – minimum string that should be annotated (the core of the phrase), for example "Bush" in "President George Bush". This tag can be empty (min=0) if there is no meaningful core of a phrase to be annotated.
- **enumeration**: the first two digits are the number of the real-world entity that the NP(s) refer to. The last two digits enumerate the mentions of this entity in the text, starting with zero. For example:
  - n-0100 = entity 1, instance 0
  - n-0101 = entity 1, instance 1
  - n-1000 = entity 10, instance 0
  - n-1001 = entity 10, instance 1

5.2.2 How much to annotate

The following examples, taken from the MUC-7 coreference task definition, illustrate what components of a phrase should be marked coreferential:

Examples with empty *min* attribute:

(25) **<n-0100 min=0>**TransCanada PipeLines Ltd.**</n-0100>** said **<npro-0101>**it**</npro-0101>** plans to shift **<npro-0102>**its**</npro-0102>** headquarters to Calgary, Alberta, from Toronto next year.

(26) **<n-0100 min=0>**London**</n-0100>** … **<n-0101 >**London-based**</n-0101>**

(27) In a report issued January 5, **<n-0100 min=0>**1995**</n-0100>**, the program manager said that there would be no new funds **<n-0101>**this year**</n-0101>**.

(28) A report was issued January 5, **<n-0100 min=0>**1995**</n-0100>**. **<n-0101>**That year**</n-0101>**, there was also…

Examples with non-empty *min* attribute:

(29) **<n-0100 min=Haden MacLellan PLC>**Haden MacLellan PLC of Surrey, England**</n-0100>**

(30) **<n-0100 min=Reuters>**Reuters Holding PLC**</n-0100>** … **<n-0101>**Reuters**</n-0101>** announced that...

(31) **<n-0100 min=contract>**The last contract**</n-0100>** you will ever get

(32) **<n-0100 min=sugar>**A large quantity of sugar**</n-0100>**

(33) **<n-0100 min=sugar>**About 200,000 tons of sugar**</n-0100>**

(34) **<n-0100 min=Ford>**Ford Motor Co. of Dearborn, Michigan**</n-0100>**

(35) **<n-0100 min=Newark>**Newark, New Jersey**</n-0100>**

(36) **<n-0100 min=December 7, 1941>**December 7, 1941, a day which will live in infamy**</n-0100>**

(37) **<n-0100 min=$1.2 million>**$1.2 million in crisp bills**</n-0100>**

(38) **<n-0100 min=shares>**20% of the shares**</n-0100>**

(39) **<n-0100 min=best>**seven of the best**</n-0100>**

(40) **<n-0100 min=five>**the five who were left standing**</n-0100>**

(41) **<n-0100 min=six>**the six youngest**</n-0100>**

(42) **<n-0100 min=rumor>**the rumor that the war had ended**</n-0100>**

(43) **<n-0100 min=Fred Frosty>**Fred Frosty, the ice cream king of Tyson's Corner**</n-0100>**

(44) **<n-0100 min=Penn Central Co.>**The Penn Central Co., which used to run a railroad**</n-0100>**

(45) **<n-0100 min=joint venture>**A joint venture with Sony**</n-0100>**

(46) **<n-0100 min=computational linguists>**Most computational linguists**</n-0100>** prefer **<npro-0101>**their**</npro-0101>** own parsers.

(47) **<n-0100 min=TV network>**Every TV network**</n-0101>** reported **<npro-0101>**its**</npro-0101>** profits yesterday. **<npro-0102>**They**</npro-0102>** plan to release full quarterly statements tomorrow.

In general, noun phrases and their modifiers should be annotated. The head noun itself should be marked as the minimal component of the phrase. Furthermore, notice that in MUC-7, Examples like (27) and (28) would not be marked as coreferential. That is because the whole date, "January 5, 1995" would be marked in the named entity task. Since what is marked in the named entity task is considered atomic, references to a component of the date, "1995", is not permissible. However, these considerations are not relevant in the present context. Therefore "1995" and "this year" in Example (27) and "1995" and "that year" in Example (28) should be marked coreferential.

6 Connectives

6.1 Basic background on connectives

Connectives are words like "because", "although", "for example", "and", "but", "or", etc. that can play an important role in determining coherence relations. Systems for establishing coherence relations such as Marcu's (2000) are mostly based on taking advantage of this fact (although Marcu's (2000) system takes into account a range of other factors as well, such as lexical relations between verbs and nouns of DSs).

It is important to note, however, that connectives are not very frequent in naturally occurring text – only about 15-20% of coherence relations in a text are usually signaled by connectives (Schauer, 2000). Furthermore, connectives can be ambiguous in a number of ways (cf. Section 6.3.1 for details).

Concerning the present study, the main point of annotating connectives and the coherence relations they signal is that it will allow studying distributional patterns of connectives. This may provide more general insights into the roles of connectives in determining coherence relations. It is possible, for instance, that connectives are preferentially used in the absence of other strong coherence cues (such as NP-bridges / anaphora or lexical relations between predicates).

6.2 Annotating connectives

Connectives are marked by XML-tags "<c> (connective) </c>". Gerunds that are clausal modifiers are marked by XML-tags "<g> (gerund) </c>", in order to make them available for further analyses. The reason for annotating these kinds of gerunds is that they are functionally similar to connectives; they can determine DS-boundaries (cf. Section 2.2 for how DS-boundaries are determined), and may signal temporal and / or causal relations between DSs (although that is an empirical question that still has to be addressed in more detail).

Examples:

(48) **<c>**but**</c>** the Senate isn't expected to act until next week at the earliest.
(49) **<g>**having succeeded**</g>**, they could move on.

6.3 Connectives and coherence relations

6.3.1 Ambiguity

As already mentioned in the previous section, connectives can be ambiguous in a number of ways:

- **Kind of coherence relation:** whereas "because" always signals a *Cause-Effect* relation, connectives like "and" and "but" can signal different coherence relations: "and" may signal almost any coherence relation (with possible exception of *Exemplification*), and "but" may signal *Contrast* or *Violated Expectation*.

- **Functional ambiguity:** connectives like "and", "but" and "or" can conjoin clauses or DSs as well as NPs (cf. Hirschberg & Litman, 1993 for possible ways of how to disambiguate these different functions). For present purposes, connectives will only be annotated if they have a DS-conjoining function, but not if they have an NP- or VP-conjoining function.

- **Ambiguity of scope:** connectives may relate two immediately adjacent utterances with each other (the DS that contains the connective and the following or the preceding DS). However, they may also connect groups of DSs (cf. examples (1) and (2) below). This scope ambiguity sometimes is resolved by punctuation, as in example (1) below.

6.3.2 Annotation of connectives and coherence relations

Coherence relations should be annotated as determined by a connective even if the kind of coherence relation is not unambiguously determined by a connective, or if the connective can have both DS- and NP- or VP-conjoining function. The idea here is that both these kinds of ambiguity – "kind of coherence relation"-ambiguity as well as functional ambiguity – can be investigated by looking at the annotation graphs. In order to investigate the "kind of coherence relation"-ambiguity, all instances of a certain connective are extracted and then investigated with respect to the kinds of coherence relations with which they co-occur (for instance, "because" should only co-occur with *Cause-Effect* relations, whereas "and" may co-occur with a range of different coherence relations). In order to investigate the functional ambiguity of connectives, instances of a connective signaling coherence relations can be contrasted with instances of that connective signaling conjoined NPs or VPs. However, with respect to the scope of coherence relations, only the scope that follows from connectives and punctuation should be annotated. Consider the following examples:

(50) Punctuation and connectives determining the full scope of a coherence relation:
(50a) "Ours is quite unique,
(50b) it's totally integrated,"
(50c) Mrs. Alptekin **said**.

(51) Connectives determining only the partial scope of a coherence relation:
(51a) robots perform specific tasks in "islands of automation,"
(51b) **but** human workers remain responsible for keeping inventory
(51c) **and** coordinating different aspects of the production line.

In Example (50) there is an a*ttribution* relation between the DS (50c) and the group consisting of DSs (50a) and (50b). The DSs (50a) and (50b) are grouped by quotation marks. This allows determining the full scope of the *atttribution* relation, including both DSs (50a) and (50b). Therefore, the complete *attribution* relation would be contained in an annotation graph for Example (50) that contains only coherence relations signaled by connectives. Notice, however, that there is also an *elaboration* relation between DS (50a) and DS (50b) that is not determined by connectives or punctuation.

Whereas in Example (50) a coherence relation between a single DS and a group of DSs is completely signaled by connectives and punctuation, this is not the case for Example (51). In Example (51) there is a *contrast* relation between the DS (51a) and the group consisting of DSs (51b) and (51c). But this relation is only signaled partially by the connective "but". The connective only signals a *contrast* relation between DS (51a) and DS (51b), but it does not signal that DS (51c) is part of the *contrast* relation as well[4].

[4] In Example (51) there is also a *parallel* relation between the DSs (51b) and (51c). This *parallel* relation is partially signaled by "and". However, additional information is necessary to unambiguously determine the *parallel* relation between DSs (51b) and (51c), in particular knowledge about the semantics of the predicates and about parallel syntactic structure between DS (51b) and DS (51c). When annotating only coherence relations that are signaled by connectives and punctuation, the fact that there is some relation between DSs (51b) and (51c) should still be annotated, however, even though connectives and punctuation alone do not signal unambiguously what kind of coherence relation it is.

The reason for this procedure is that it should be possible to investigate to what extent the scope of coherence relations is determined by connectives. This can be done by comparing the scope of a coherence relation in the complete annotation graph with the scope any given relation has in the annotation graph that only contains the coherence relations determined by connectives and punctuation.

7 Annotation tools and file formats

7.1 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 elabotation
2 <-> 4 contrast

Using the *annotator* tool (cf. Section 7.3) 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 par *// this line is in the annotation file*
3 3 1 1 par *// 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 Mengel & Lezius (2000).

7.2 Emacs

Emacs is used to annotate anaphora as well as connectives. XML-tags can be highlighted, using Emacs HTML syntax highlighting.

7.3 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 7.1)
- 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)

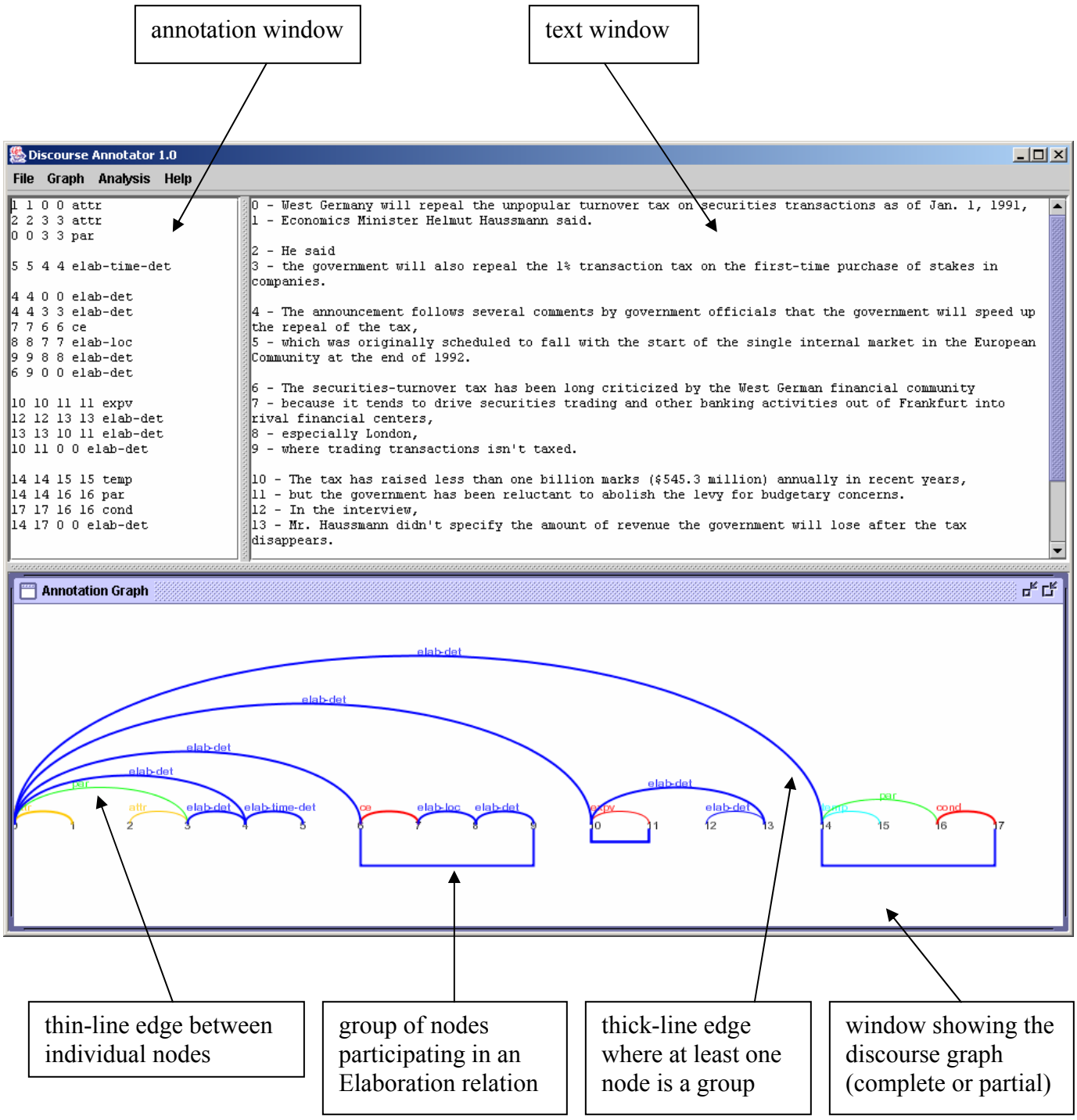


Figure 4. Screenshot of the *annotator* tool.

7.4 Perl scripts

A number of Perl scripts are available for further data processing. This section describes the most important scripts.

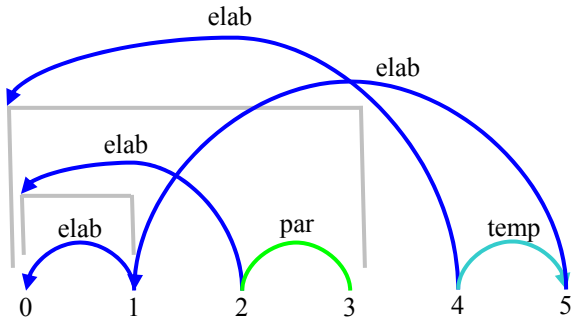
7.4.1 Perl script *annotator2hierarchical.pl*

- **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.

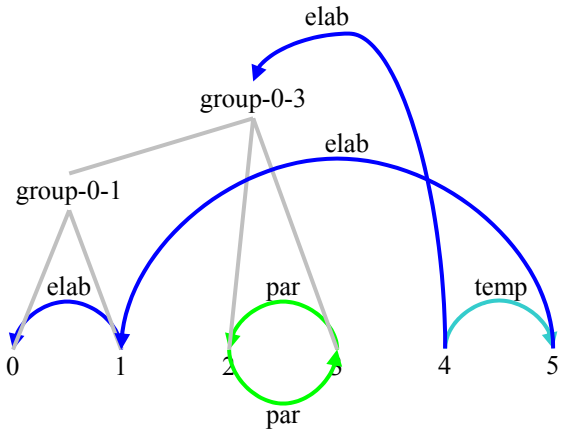
- **Input to *annotator2hierarchical.pl*:**
  - **Output of *annotator* tool:**
    1 1 0 0 elab
    3 3 2 2 par
    2 2 0 1 elab
    4 4 5 5 temp
    4 4 0 3 elab
    5 5 1 1 elab

  - **Graphical representation in *annotator* tool:**



- **Output of *annotator2hierarchical.pl*:**
  - **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

  - **Graphical representation:**



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*.

7.4.2 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).

7.5 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*

8 References

[1] Barzilay, R (1997). *Lexical chains for summarization*. MSc Thesis, Department of Mathematics & Computer Science, Ben-Gurion University, Beersheva, Israel.

[2] Buntine, W (1994). *Operations for learning with graphical models*. Manuscript, NASA Ames Research Center, Moffett Field, CA.

[3] Carletta, J (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2): 249-254.

[4] Carlson, L, Marcu, D & Okurowski, ME (2002). *RST Discourse Treebank*. Philadelphia, PA: Linguistic Data Consortium.

[5] Cristea, D, Ide, N & Romary, L (1998). Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pp. 281-285. Montreal, Canada, August 1998.

[6] Hirschberg, J & Grosz, BJ (1992). Intonational features of local and global discourse structure. In *Proceedings of the Speech and Natural Language Workshop*, pp. 441--446, Harriman, New York.

[7] Hirschberg, J & Litman, D (1993). *Empirical studies on the disambiguation of cue phrases*. Manuscript, AT&T Bell Laboratories, Murray Hill, NJ.

[8] Hobbs, JR (1985). *On the coherence and structure of discourse*. Technical Report CSLI-85-37, CSLI, Palo Alto, CA.

[9] Hobbs, JR, Stickel, M, Appelt, D & Martin, P (1990). Interpretation as abduction. Manuscript, AI Center, SRI International, Menlo Park, CA.

[10] Hovy, EH & Maier, E (1998). *Parsimonious or profligate: How many and which discourse structure relations?* Manuscript, Information Sciences Institute, USC, Los Angeles, CA.

[11] Huang, C & Darwiche, A (1994). Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning, 11*: 1-158.

[12] Jurafsky, D & Martin, JH (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall, Upper Saddle River, NJ.

[13] Kehler, A (1995). *Interpreting cohesive forms in the context of discourse inference*. PhD thesis, Department of Computer Science, Harvard University, Cambridge, MA.

[14] Kehler, A (2000). Coherence and the resolution of ellipsis. *Linguistics and Philosophy, 23*: 533-575.

[15] Kehler, A (2002). *Coherence, reference, and the theory of grammar*. CSLI Publications, Stanford, CA.

[16] Mann, WC & Thompson, SA (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3): 243-281.

[17] Marcu, D (2000). *The theory and practice of discourse parsing and summarization*. Cambridge, MA: MIT Press.

[18] Mengel, A & Lezius, W (2000). An XML-based representation format for syntactically annotated corpora. In: *Proceedings of the LREC 2000*. Athens, Greece.

[19] Moore, JD & Pollack, ME (1992). *A problem for RST: The need for multi-level discourse analysis*. Manuscript, Dept. of Computer Science, University of Pittsburgh, PA.

[20] Polanyi, L & van den Berg, M (1997). *Discourse structure and discourse interpretation*. Manuscript.

[21] Schauer, H (2000). *Referential structure and coherence structure*. Conference TALN 2000, Lausanne, Switzerland.

[22] Strube, M & Hahn, U (1995). *ParseTalk about textual ellipsis*. Manuscript, Computational Linguistics Research Group, Freiburg University, Germany.

[23] Webber, B, Knott, A, Stone, M & Joshi, A (1999). *Discourse relations: A structural and presuppositional account using Lexicalised TAG*. Manuscript, University` of Edinburgh.

[24] Webber, B, Stone, M, Joshi, A & Knott, A (submitted). Anaphora and discourse deictics. *Computational Linguistics*.

[25] Wiebe, JM (1994). *Issues in linguistic segmentation*. Manuscript, Dept. of Computer Science, New Mexico State University, NM.

[26] Wolf, F, Gibson, E & Desmet, T (2003). *Pronoun processing and coherence*. Poster presented at the 16th Annual CUNY Conference on Human Sentence Processing, MIT, Cambridge, MA.

[27] Wolf, F & Gibson, E (2003). *The descriptional adequacy of trees for representing discourse coherence*. Manuscript, MIT, Cambridge, MA.

[28] MUC-7: http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/muc_7_toc.html

9 Appendix A –Annotation procedure "recipes"

9.1 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:

Coherence Relation	Connective
Cause-Effect	because
Violated Expectation	although
Condition	if…then
Parallel	(and) similarly
Contrast	by contrast
Temporal Sequence	and then
Attribution	according to…
Example	for example
Elaboration	also, furthermore, in addition
Generalization	in general

9.2 Important distinctions

- **Difference *Example – Elaboration*:** an *Example* sets up an additional entity (the example), whereas an E*laboration* 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.

9.3 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.