Discourse Tagging Reference Manual

Lynn Carlson and Daniel Marcu

We would like to acknowledge the annotators who participated in the tagging of this corpus and previous work leading up to this effort. Their contributions to the ideas presented here are significant: Estibaliz Amorrortu, Jean Hobbs, John Kovarik, Toby Merriken, Mary Ellen Okurowski, Norb Rieg, Magdalena Romera, Maki Watanabe, and Markell West.

1.0 Introduction

This reference manual presents the guidelines used to develop a large discourse-annotated corpus for community-wide use. The resulting resource consists of 385 documents of American English selected from the Penn Treebank (Marcus, et al, 1993), annotated in the framework of Rhetorical Structure Theory.[1] We assume here that the reader is familiar with the basic principle of RST, as presented in Mann and Thompson (1988). We also refer the reader to Carlson et al. (2001), which describes a number of issues and challenges in building this corpus, and to Marcu et al. (1999), which addresses experimental issues in annotating the discourse structure of entire texts in the RST framework.

Fundamentally, the guidelines presented here follow those outlined in *Instructions for Manually Annotating the Discourse Structure of Texts* (Marcu, 1999). However, this manual contains further refinements to those rules, reflecting the nature of the texts -- all of which are taken from the Wall Street Journal -- and motivated by the need to present as consistent a set of tagged documents as possible for use by the research community. We have included heuristics and diagnostics describing how we made fine-grained decisions on tagging these documents, using examples from the tagged corpus and additional WSJ articles from the Treebank. Included are discussions on text segmentation, embedded discourse units, attribution verbs, cue phrases, and selecting relations.

The manual includes a number of appendices which are intended as quick references to aid the discourse taggers.

Some formatting conventions used for the examples in this manual are:

• All examples are shown in `Courier` font to distinguish them from the body of the text. Elementary discourse units are marked in square brackets; the source of the example (wsj_doc#) is shown as a subscript at the end of the example:

(1) `[Wall Street is understandably skeptical.]`$_{\texttt{wsj\_0604}}$

• Embedded discourse units, such as parentheticals and relative clauses, are shown in small font:

(2) `[Mr. Volk, 55 years old, succeeds Duncan Dwight,]`
`[who retired in September.]`$_{\texttt{wsj\_0600}}$

1. For information on obtaining the corpus and accompanying documents see: http://www.isi.edu/~marcu/discourse.

• When discussing relations, the nucleus is shown in normal font and the satellite is shown in italics:

(3) [Only a few months ago, the 124-year-old securities firm seemed to be on the verge of a meltdown,] [*racked by internal squabbles and defections.*]$_{wsj\_0604}$

• When a particular issue is in focus, all segmentation will be shown, but the unit or text fragment relevant to the issue being discussed will be underlined for clarity. Boldface may be used to highlight particular lexical or syntactic cues that are relevant to determining the discourse structure. Superscripts at the end of a bracketed unit mark the unit number. For example, the sentence below contains four EDUs. However, since the focus of the section is on non-finite postnominal modifiers, only units [2] and [4] are underlined, and the non-finite verb forms are shown in boldface:

(4) [The results underscore Sears's difficulties^{1}] [**in implementing** the "everyday low pricing" strategy^{2}] [that it adopted in March, as part of a broad attempt^{3}] [**to revive** its retailing business.^{4}]$_{wsj\_1105}$

• Further distinctions in the examples may be made with italics or double underlined, and will be noted accordingly.

2.0 Segmenting Texts into Elementary Units

The first step in characterizing the discourse structure of a text in our protocol is to determine the elementary discourse units (EDUs), which are the minimal building blocks of a discourse tree. Mann and Thompson (1988, p. 244) state that "RST provides a general way to describe the relations among clauses in a text, whether or not they are grammatically or lexically signalled." Yet, applying this intuitive notion to the task of producing a large, consistently annotated corpus is extremely difficult, because the boundary between discourse and syntax can be very blurry. The examples below, which range from two distinct sentences to a single clause, all convey essentially the same meaning, packaged in different ways:

(5) [Xerox Corp.'s third-quarter net income grew 6.2% on 7.3% higher revenue.] [This earned mixed reviews from Wall Street analysts.]

(6) [Xerox Corp's third-quarter net income grew 6.2% on 7.3% higher revenue,] [which earned mixed reviews from Wall Street analysts.]

(7) [Xerox Corp's third-quarter net income grew 6.2% on 7.3% higher revenue,] [earning mixed reviews from Wall Street analysts.]$_{wsj\_1109}$

```
 (8)  [The 6.2% growth of Xerox Corp.’s third-quarter
net income on 7.3% higher revenue earned mixed reviews
from Wall Street analysts.]
```

In Example 5, there is a consequential relation between the first and second sentences. Ideally, we would like to capture that kind of rhetorical information regardless of the syntactic form in which it is conveyed. However, as examples 6-8 illustrate, separating rhetorical from syntactic analysis is not always easy. It is inevitable that any decision on how to bracket elementary discourse units necessarily involves some compromises.

Researchers in the field have proposed a number of competing hypotheses about what constitutes an elementary discourse unit. While some take the elementary units to be clauses (Grimes, 1975; Givon, 1983; Longacre, 1983), others take them to be prosodic units (Hirschberg and Litman, 1993), turns of talk (Sacks, 1974), sentences (Polanyi, 1988), intentionally defined discourse segments (Grosz and Sidner, 1986), or the “contextually indexed representation of information conveyed by a semiotic gesture, asserting a single state of affairs or partial state of affairs in a discourse world,” (Polanyi, 1996, p.5). Regardless of their theoretical stance, all agree that the elementary discourse units are non-overlapping spans of text.

Our goal was to find a balance between granularity of tagging and ability to identify units consistently on a large scale. In the end, we chose the clause as the elementary unit of discourse, using lexical and syntactic clues to help determine boundaries. A few refinements to this basic principle are enumerated below, with reference to the section of the manual that discusses the phenomenon in more detail.

• Clauses that are subjects or objects of a main verb are not treated as EDUs. (Section 2.2)

• Clauses that are complements of a main verb are not treated as EDUs. (Section 2.3)

• Complements of attribution verbs (speech acts and other cognitive acts) are treated as EDUs. (Section 2.4)

• Relative clauses, nominal postmodifiers, or clauses that break up other legitimate EDUs, are treated as embedded discourse units. (Section 2.9)

• Phrases that begin with a strong discourse marker, such as *because, in spite of, as a result of, according to,* are treated as EDUs. (Section 2.10)

A number of other phenomena that caused difficulty in segmentation are addressed as well. These include: coordination (Section 2.5), syntactic focusing devices (Section 2.6), temporal expressions (Section 2.7), correlative subordinators (Section 2.8), and punctuation (Section 2.11). Appendix I presents a table summarizing the syntactic phenomena we addressed, and whether or not they trigger an EDU boundary.

In making our final decisions regarding segmentation, our overriding concern was for consistency in tagging across a large number of texts. This meant a number of compromises had to be made, including occasionally sacrificing some potentially discourse-relevant information.

2.1 Basic Segmentation Principles

The basic elementary discourse unit (EDU) is a clause. (A limited number of phrasal EDUs are permitted as well; see Section 2.10 below for details.) Some clear-cut examples of sentences containing two clausal EDUs are enumerated below -- each of these has a superordinate clause, and a subordinate clause, containing a discourse marker:

(9) [Such trappings suggest a glorious past] [**but** give no hint of a troubled present.]$_{wsj\_1302}$

(10) [**Although** Mr. Freeman is retiring,] [he will continue to work as a consultant for American Express on a project basis.]$_{wsj\_1317}$

(11) [Share prices in Frankfurt closed narrowly mixed] [**after** Wall Street opened stronger.]$_{wsj\_0374}$

(12) [Previously, airlines were limiting the programs] [**because** they were becoming too expensive.]$_{wsj\_1192}$

(13) [Sears, Roebuck & Co. is struggling] [**as** it enters the critical Christmas season.]$_{wsj\_1105}$

(14) [For years, the five New York exchanges have been talking about cooperating in various aspects of their business] [**in order to** improve the efficiency of their operations.]$_{wsj\_0664}$

Note that for the EDU which is the subordinate clause, the discourse cue may be preceded by an adverb:

(15) [More people are remaining independent longer] [presumably **because** they are better off physically and financially.]$_{wsj\_2366}$

(16) [Thus, the very fickleness of baby boomers may make it possible to win them back,] [just **as** it was possible to lose them.]$_{wsj\_1377}$

In other cases, the subordinate clause may not contain a lexical discourse cue. For example, a participial clause may form an EDU:

(17) [Xerox Corp.’s third-quarter net income grew 6.2% on 7.3% higher revenue,] [**earning** mixed reviews from Wall Street analysts.]$_{wsj\_1109}$

(18) [The Singapore and Kuala Lumpur stock exchanges are bracing for a turbulent separation,] [**following**

Malaysian Finance Minister Daim Zainuddin's long-awaited announcement] [that the exchanges will sever ties.]$_{wsj\_0613}$

An infinitival modifier may also be an EDU, provided that it is not a complement of the verb (see section 2.3.1 for discussion):

(19) [-- and that it would quickly make its way to the Supreme Court] [**to be** ultimately **resolved.**]$_{wsj\_0609}$

(20) [The group's president, Peter Christanthopoulos, wasn't in his office Friday afternoon] [**to comment**.]$_{wsj\_2315}$

2.2 Clausal Subjects and Objects

Clausal subjects and objects of verbs should not be treated as elementary discourse units:

(21) [Deciding what constitutes "terrorism" can be a legalistic exercise.]$_{wsj\_1101}$

(22) [Atco Ltd. said its utilities arm is considering building new electric power plants, ...]$_{wsj\_2309}$

(23) [Making computers smaller often means sacrificing memory.]$_{wsj\_2387}$

Similarly, clausal objects of prepositional phrases should not be treated as elementary discourse units:

(24) [So far, it appears cautious about taking the big step.]$_{wsj\_0634}$

(25) [We have opened eyes to being a little less conservative and more imaginative in how to present the news.]$_{wsj\_0633}$

(26) [The bank says it's interested in lending to individuals and small and medium-sized companies.]$_{wsj\_0616}$

However, an entire prepositional phrase with a clausal object may be an EDU. For example, in (27) below, the prepositional phrase is the satellite of a CIRCUMSTANCE relation, while in (28), the prepositional phrase is the satellite of a MEANS relation:

(27) [Canadian Utilities isn't alone] [**in** exploring power generation opportunities in Britain, ...]$_{wsj\_2309}$

(28) [Maybe she could step across the plaza to the Met [-- where she still has to make her debut --] and help out her Czech compatriot] [**by** singing the slow parts of "Traviata."]$_{wsj\_1154}$

2.3 Clausal Complements

Clausal complements of verbs are normally not broken out into separate EDUs. We make an exception to this in the case of verbs of attribution (see Section 2.4). Below are some types of clausal complements that should be treated as one unit with the main verb. In the examples, the main verb is shown in italics, and the verbal complement is underlined, with the verb form in boldface:

2.3.1 Infinitival Complements

Never break infinitival complements of verbs as separate EDUs. (See additional discussion on infinitival complements of attribution verbs in Section 2.4.)

(29) [He said] [the thrift will *try* **to get** regulators **to reverse** the decision.]$_{wsj\_2360}$

(30) [Ideally, we'd *like* **to be** the operator {of the project} and a modest equity investor.]$_{wsj\_2309}$

(31) [Even the Soviet Union *has* Peter the Great **to rediscover**,] [should it choose to.]$_{wsj\_1929}$

However, an infinitival complement should not be confused with an infinitival clause that functions as the satellite of a PURPOSE relation, which can be tested by substituting the phrase *in order to* for the *to* clause. The following example has two infinitival clauses: the first is a complement (underlined) and the second, a purpose clause (double underlined):

(32) [A grand jury has been investigating whether officials at Southern Co. *conspired* **to cover up** their accounting for spare parts] [**to evade** federal income taxes.]$_{wsj\_0619}$

In some situations, determining whether a *to*-infinitive is a complement to a main verb or a purpose clause can be tricky: We list some examples below of cases which we treated as complements:

(33) [Since its premiere Sept. 16, the show] [on which Ms. Chung appears] [has *used* an actor **to portray** the Rev. Vernon Johns, a civil-rights leader, and one **to play** a teenage drug dealer.]$_{wsj\_0633}$

(34) [With the golden share as protection, Jaguar officials have rebuffed Ford's overtures, and *moved* instead **to forge** an alliance with GM.]$_{wsj\_0632}$

(35) [The network deals a lot with unknowns,] [including Scott Wentworth,] [who portrayed Mr. Anderson,] [and Bill Alton as Father Jenco,] [but the network *has* some big names **to contend** with, too.]$_{wsj\_0633}$

(36) ["We *have tried* our best **to tell** the people in Bataan] [that maybe this time it will not go to them,] [but certainly we *will do* our best **to encourage** other investors to go to their province,"] [Mrs. Aquino told Manila-based foreign correspondents.]$_{wsj\_0606}$

2.3.2 Participial Complements

Participial complements of verbs should not be treated as separate EDUs:

(37) [Last March, this newspaper reported on widespread allegations] [that the company *misled* many customers **into purchasing** more credit-data services] [than needed.]$_{wsj\_1157}$

(38) [The company's current management *found* itself "**locked** into this," he said.]$_{wsj\_1103}$

(39) [Insurers *could see* claims **totaling** nearly $1 billion from the San Francisco earthquake,] far less than the $4 billion from Hurricane Hugo.]$_{wsj_0675}$

(40) [Bowing to criticism,] [Bear Stearns, Morgan Stanley and Oppenheimer *joined* PaineWebber in **suspending** stock-index arbitrage trading for their own accounts.]$_{wsj\_0675}$

2.4 Attribution Verbs

Normally, clausal complements of verbs are not considered to be EDUs. We make exception to this in the case of clausal complements of *attribution verbs,* including both speech acts and other cognitive acts.

2.4.1 Reported Speech

Speech acts -- verbs that are used to report both direct and indirect speech -- should be segmented and marked for the rhetorical relation of ATTRIBUTION, if the following two conditions are met: 1) there is an explicit source of the attribution, and 2) the attribution predicate takes a clausal complement that is not infinitival. Mark the source of the attribution (the clause containing the report-

ing verb) as the satellite, and the content of the reported message (which must be in a separate clause) as the nucleus:

(41) *[Mercedes officials said]* [they expect flat sales next year] [even though they see the U.S. luxury-car market expanded slightly.]$_{wsj\_1196}$

(42) *[Bush indicated]* [there might be "room for flexibility" in a bill] [to allow federal funding of abortions for poor women] [who are victims of rape and incest.]$_{wsj\_2356}$

(43) *[The legendary GM chairman declared]* [that his company would make "a car for every purse and purpose."]$_{wsj\_1377}$

The following are typical attribution verbs, whose complements are generally treated as separate EDUs:

say, tell, state, announce, declare, suggest, advise, report, indicate, point out, explain, ask

We make an exception to this rule in the case of infinitival complements. In such cases, the main verb and the subordinate clause are treated as one unit:

(44) [The former first lady of the Philippines asked a federal court in Manhattan **to dismiss** an indictment against her...]$_{wsj\_0617}$

(45) [Scott has spoken to his attorney,] [who has advised him **not to talk** to anybody.]$_{wsj\_0335}$

When the source of the attribution is a phrase, beginning with the expression *according to*, rather than a clause, it should also be marked as an elementary discourse unit:

(46) [The shares represented 66% of his Dun & Bradstreet holdings,] [**according to** the company.]$_{wsj\_1157}$

If the clause containing the reporting verb does not specify the source of the attribution, such is in passive voice constructions, or generic expressions like *it is said*, then a relation of attribution does not hold, and the reporting and reported clauses are treated as one unit:

(47) [Earlier yesterday, the Societe de Bourses Francaises was told that a unit of Framatome S.A. also bought Navigation Mixte shares, this purchase covering more than 160,000 share.]$_{wsj\_0340}$

If the clause containing the reporting verb does not specify the source in that clause, *but the source can be identified elsewhere in the sentence, or the nearby context,* as in the following examples, a relation of ATTRIBUTION still is considered to hold (italics = attribution satellite; boldface = attribution source):

(48) [Soon after the merger, moreover, Federal's management asked Tiger's pilots to sign **an agreement**] *[stating]* [that they could be fired any time,] [without cause or notice.]$_{wsj\_1394}$

(49) [**Shearson Lehman Hutton** gave small investors some welcome news] *[by announcing]* [that it would no longer handle index-arbitrage-related program trades for its accounts.]$_{wsj\_0327}$

2.4.2 Cognitive Predicates

Cognitive predicates, including verbs that express feelings, thoughts, hopes, etc., should also be segmented and marked for the rhetorical relation of ATTRIBUTION, according to the two conditions described above for reporting verbs:

(50) *[Analysts estimated]* [that sales at U.S. stores declined in the quarter, too.]$_{wsj\_1105}$

(51) *["I don't know]* [whether it was done properly or not,] [because I'm not a lawyer,"] [he said in a telephone interview yesterday.]$_{wsj\_1366}$

The following cognitive predicates should generally be marked for the relation of ATTRIBUTION, if the conditions mentioned above hold:

think, believe, know, imagine, suppose, conjecture, wish, hope, predict, fear, estimate, calculate, anticipate, expect, dream

However, do not segment these cognitive predicates if they are followed by an infinitival complement, such as the following:

try to, attempt to, appear to, seem to, seek to, aim to, endeavor to, ask to, expect to, wish to, hope to, want to, decide to, learn to, decline to.

Some examples from the corpus:

(52) [Elcotel Inc. *expects* fiscal second-quarter earnings **to trail** 1988 results, ...]$_{wsj\_2317}$

(53) [Israel *wants* **to end** the dialogue,...]$_{wsj\_1101}$

When an attribution predicate takes a clausal complement that begins with an interrogative pronoun, the complement should be segmented as a separate EDU, as long as the complement is not infinitival:

(54) *[One cannot imagine]* [**how** you live] [when you live those double and triple lives.]$_{wsj\_1367}$

(55) [A spokesman for GM, the No. 1 auto maker, declined to say] [**how many** Jaguar shares that company owns.]$_{wsj\_0633}$

(56) [Harsco declined to say] [**what** country placed the order.]$_{wsj\_1395}$

(57) [“People are very concerned about] [**who** is going to step up to the plate and buy municipal bonds in the absence of institutional buyers.”]$_{wsj\_0671}$

If the interrogative pronoun is followed by an infinitival complement, then the clause is not segmented and the ATTRIBUTION relation does not apply:

(58) [“We know **how** to get from capitalism to socialism,”] [Sergiusz Niciporuk is saying one afternoon.] [“We don't know **how** to get from socialism to capitalism.”]$_{wsj\_1146}$

Clauses containing verbs of perception, when used in the manner of a cognitive predicate, such as *I see that*, should be marked as the satellite of an ATTRIBUTION relation, when the source of the attribution is provided. Verbs in this category include *see, feel, hear, sense.*

(59) [We *hear*] [that HUD Secretary Jack Kemp is toying with going along with some of the Cranston-Mitchell proposals.]$_{wsj\_0309}$

(60) [Thirty-five percent attend religious services regularly;] [at the same time, 60% *feel*] [that in life one sometimes has to compromise one’s principles.]$_{wsj\_2366}$

However, when the clause containing the perception verb does not specify a source, the ATTRIBUTION relation does not apply and the unit should not be segmented from the *that* clause. This is typical in constructions like with verbs like *seem, look,* or *appear* (with *it* as subject), and in passive constructions:

(61) [It wasn’t known to what extent, if any, the facility was damaged.]$_{wsj\_1915}$

(62) [It is hoped that other Japanese would then follow the leader.]$_{wsj\_0300}$

(63) [..., but recently, it appears Sotheby’s has been returning the compliment.]$_{wsj\_1928}$

An exception is made if the source is indicated with one of the above constructions, as in:

(64) [“It *seems* **to me**] [that a story like this breaks just before every important Cocom meeting,”] [said a Washington lobbyist for a number of computer companies.]$_{wsj\_2326}$

2.5 Coordination

Coordinated sentences and clauses are broken into separate EDUs, while coordinated verb phrases are not. The examples below illustrate varying degrees of coordination:

2.5.1 Coordinated Sentences

Coordinated sentences are marked as elementary discourse units. The conjoins may be separated by a comma, followed by a coordinating conjunction; a semi-colon; colon; or other device. Some examples of coordinate sentences are given below, with the boundaries of the conjoins marked as EDUs:

(65) [Inventories are creeping up;] [car inventories are already high,] [**and** big auto makers are idling plants.]$_{wsj\_0359}$

(66) [Declining issues outnumber advancers 551 to 349;] [224 issues were unchanged.]$_{wsj\_0374}$

(67) [And some carriers are facing other unexpected headaches:] [USAir, for example, blamed some of its loss on merger expenses and on disruptions] [caused by Hurricane Hugo last month.]$_{wsj\_1192}$

A coordinated sentence may have a gapped verb, as in unit [3] below:

(68) [Some gripes are about red tape are predictable:^{1}] [Architects complain about a host of building regulations,^{2}] [auto leasing companies about car insurance rules.^{3}]$_{wsj\_1162}$

2.5.2 Coordinated Clauses

Coordinated clauses differ from coordinated sentences in that while the conjoins each contain a distinct verb phrase, they may share an ellipted subject, and may also share an ellipted auxiliary verb and adverb.

2.5.2.1 Coordination in Superordinate Clauses

When coordination clauses occur as superordinate clauses, they are treated like coordinated sentences, and should be marked as EDUs. The ellipted components of the examples below are shown in boldface:

(69) [**The issue was** oversubscribed] [and “doing very well,”] [according to an official with lead underwriter Morgan Stanley.]$_{wsj\_1322}$

(70) [**Stock prices,** meanwhile, posted significant gains in later trading,] [and closed down by only 3.69 points on the day.]$_{wsj\_1102}$

In a limited number of cases, the main verb may even be ellipted; however, this is only allowed when there are strong rhetorical cues marking the discourse structure:

(71) [Back then, Mr. Pinter was **not only** the angry young playwright,] [**but also** the first] [to use silence and sentence fragments and menacing stares, almost to the exclusion] [of what we previously understood to be theatrical dialog.]$_{wsj\_1936}$

Clausal conjoins should be segmented even when they are preceded or followed by a word or phrase which appears to scope over both conjoins. If the word or phrase precedes the conjoins, include it with the first one; otherwise, include it with the last:

(72) [A spokesman for the New York-based food and tobacco giant,] [taken private earlier this year in a $25 billion leveraged buy-out by Kohlberg Kravis Roberts & Co.,] [confirmed] [that it is shutting down the RJR Nabisco Broadcast unit,] [and dismissing its 14 employees, **in a move**] [**to save money**.]$_{wsj_2315}$

2.5.2.2 Coordination in Subordinate Clauses

The examples above all illustrate coordinated clauses as main or independent clauses in a sentence. When coordinated clauses are subordinated to a main verb, they may or may not be segmented as separate EDUs. The EDU status of such constructions depends on whether or not the subordinate construction would normally be segmented as an EDU if it were a single clause, rather than a number of coordinated clauses.

For example, in (73) below the series of clauses beginning with *so that their clients can...* are segmented as separate EDUs which form the satellite of a PURPOSE relation. The satellite consists of a multinuclear LIST of coordinated clauses:

(73) Equipped with cellular phones, laptop computers, calculators and a pack of blank checks,] [they parcel out money] *[so that **their clients can** find temporary living quarters,] [buy food,] [replace lost clothing,] [repair broken water heaters,] [and replaster walls.]*$_{file3}$

The discourse sub-tree for this sentence is as follows:

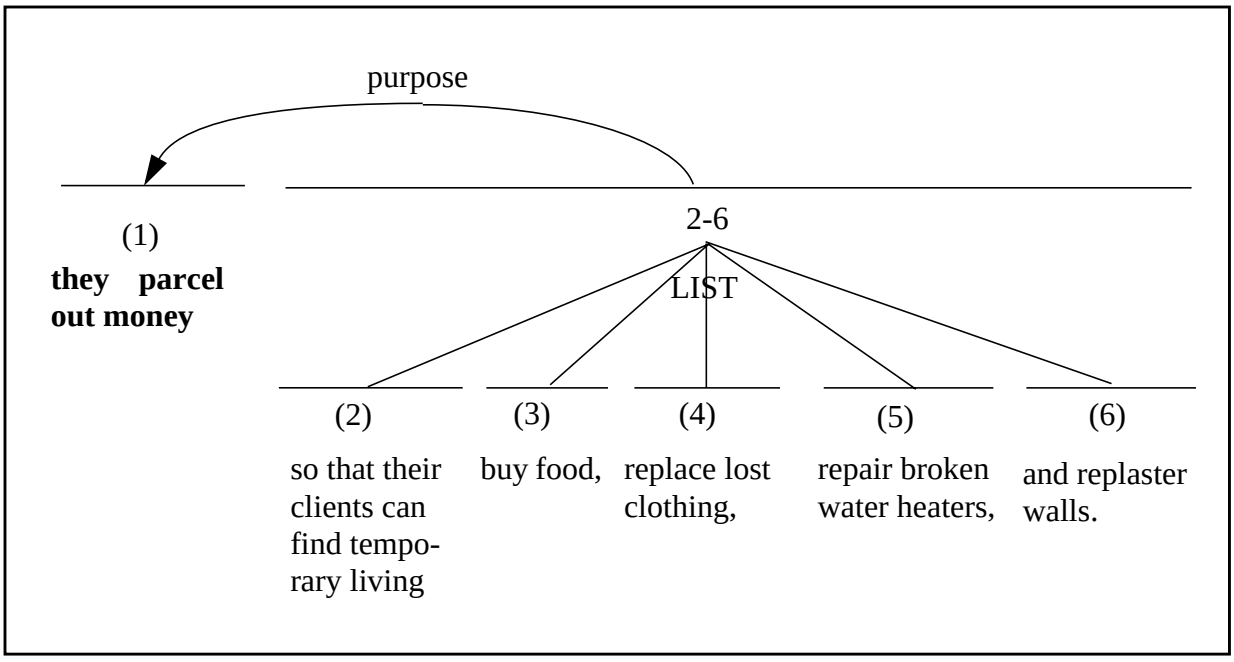

Figure 1: Coordinated Clauses as EDUs in Subordinate Constructions

However, if we rephrase this example to read:

(74) [They need to find temporary living quarters, buy food, replace lost clothing, repair broken water heaters, and replaster walls.]$_{file3}$

the situation is different. According to our rules, infinitival complements are never EDUs. Therefore, even though the infinitival complement of *need* contains multiple clauses, these would not be segmented. The entire sentence is one EDU.

Compare the examples below. In (75), the two coordinated clauses are segmented as a two-part LIST, which constitutes the satellite of a PURPOSE relation with respect to the main clause. In (76), on the contrary, the coordinated clauses are not segmented, since the verb phrases *restructure financially* and *improve its balance sheet* are (bare) infinitival complements of the verb *help*:

(75) [Some of the funds will be used] *[to demolish unstable buildings] [and clear sites for future construction.]*$_{file3}$

(76) [NBI also said] [it has hired Prudential-Bache Securities Inc. as its financial adviser and investment banker] [**to help** it restructure financially and improve its balance sheet.]$_{wsj\_0647}$

Additional examples of coordinated clauses that are segmented in subordinate constructions are listed below. In Example (77), the coordinated clauses form the satellite of a MANNER relation with respect to the main clause, while in Example (78), the coordinated clauses form the satellite of a PURPOSE relation with respect to the preceding clause:

(77) [“And you didn’t want me to buy earthquake insurance,” says Mrs. Hammack,] *[reaching across the table] [and gently tapping his hand.]*$_{\text{file3}}$

(78) [The SEC has alleged] [that Mr. Antar aimed to pump up the company’s stock price through false financial statements] *[in order to sell his stake] [and reap huge profits.]*$_{\text{file4}}$

Finally, we show an example with both segmented and non-segmented infinitival clauses. The coordinated infinitival clauses in unit [1] are not segmented, because they are complements of the verb *urge*, while the infinitival clause in unit [3] is segmented as a postmodifier of the noun phrase *the power* (see Section 2.9.2 for further discussion):

(79) [But some analysts, particularly conservative legal scholars, have urged Mr. Bush not to wait for explicit authorization but simply to assert^{1}] [that the Constitution already implicitly gives him the power^{2}] [to exercise a line-item veto.^{3}]$_{\text{wsj\_1133}}$

For examples of coordinate clauses as postmodifiers of nominal constructions, see the discussion on Embedded Discourse Units, Section 2.9 below.

2.5.3 Coordinated Verb Phrases

Coordinated verb phrases should not be broken into separate EDUs. One test for whether you have coordinated verb phrases rather than coordinated clauses is when two transitive verbs (underlined) share the same direct object (double underlined), as in the examples below:

(80) [Once inside, she spends fours hours measuring and diagramming each room in the 80-year-old house,...]$_{\text{file3}}$

(81) [Under Superfund, those] [who owned, generated or transported hazardous waste] [are liable for its cleanup,] [regardless of whether their actions were legal at the time.]$_{\text{wsj\_1331}}$

An example of coordinated verb phrases with intransitive verbs is given below. In this example, the verbs share an adjectival modifier, *overseas*:

(82) [Today,] [though the public is barely aware,] [much of U.S. industry, particularly companies] [manufacturing or selling overseas,] have made metrics routine.]$_{wsj\_0676}$

2.6 Syntactic Focusing Devices

There are a number of syntactic focusing devices that have the effect of dividing a single sentence into two separate clauses, in order to provide emphasis on a particular element in the sentence. They include cleft, extraposition, and pseudo-cleft constructions, discussed below. Such constructions are treated as single EDUs.

2.6.1 Cleft

The cleft construction brings focus to an element by moving it to the front of the sentence (see Quirk and Greenbaum, 1973, pp. 414-416). This device divides a single clause into two clauses, the first of which begins with *it* plus a form of the verb *be*, followed by the fronted element (underlined in the examples). Even though the result is two distinct clauses, we regard the two parts of a cleft sentence as a single EDU:

(83) [It was not until the early 1970s that Cambridge Prof. Harry Whittington and two sharp graduate students began to publish a reinterpretation of the Burgess Shale.]$_{wsj\_1158}$

(84) [It was just the culture of the industry that kept it from happening.]$_{wsj\_0317}$

(85) [It was the Senate Republicans, though, who had edged away from the veto strategy.]$_{wsj1939}$

2.6.2 Extraposition

Extraposition is a linguistic device that removes an element from its normal position and places it towards the end of the sentence, leaving a substitute form (such as *it*) in its place. (Quirk and Greenbaum, 1973, pp. 422-423). The effect of the postponement is to achieve end-focus and end-weight. Extraposition is the more standard word order when a clausal subject is involved, as in the examples below. As with the cleft construction, the result of extraposition is to create two distinct clauses out of one. Again, we treat this phenomenon as a single EDU:

(86) [It is difficult to analyze how much of their anger was due to Recruit, the sex scandals, or the one-yen coins in their purses,...]$_{wsj\_1120}$

(87) [It's hard to imagine how the markets were speculating,...]$_{wsj\_1932}$

(88) [It is insulting and demeaning to say that scientists "needed new crises to generate new grants and

contracts" and that environmental groups need them to stay in business.]$_{wsj\_0360}$

(89) [Both women say] [they find it distasteful that CBS news is apparently concentrating on Mr. Hoffman's problems as a manic-depressive.]$_{wsj\_0633}$

Constructions with *it seems/appears* are similar to clausal extraposition, and also will be treated as single EDUs (unless they occur in expressions where an attribution source is mentioned, e.g., *it seems to me*; see Section 2.4 for details).

(90) [It appears that the only thing Congress is learning from the HUD story is how to enlarge its control of the honey pot] [going to special interests.]$_{wsj\_0309}$

2.6.3 Pseudo-cleft

Similar to the cleft construction, pseudo-cleft is a focusing device that divides a single sentence into two parts, using a *wh*-form. As with cleft and extraposition, the two resulting parts are treated as a single EDU:

(91) [What defeated General Aoun was not only the weight of the Syrian army.]$_{wsj\_1141}$

(92) ["Kodak understands] [HDTV is where everyone is going,"] [says RIT's Mr. Spaull."]$_{wsj\_1386}$

2.7 Temporal Expressions

Temporal expressions are marked off as separate EDUs if they occur in clausal constructions:

(93) [The dollar finished lower yesterday,] [**after** tracking another rollercoaster session on Wall Street.]$_{wsj\_1102}$

(94) [House and Senate appropriators sought to establish a Nov. 30 deadline] [**after** which their bill would become the last word] [on how funds are distributed.]$_{wsj\_0328}$

(95) [It was the market's biggest gain] [**since** rising more than 7 points on Oct. 19.]$_{wsj\_0327}$

(96) [Navigation Mixte's chairman had suggested] [that friendly institutions were likely to buy its stock] [**as soon as** trading opened Monday.]$_{wsj\_0340}$

Temporal clauses may contain a subordinating conjunction (*before, after*), which may be preceded by one or more modifiers, which are typically included as part of the EDU. Several types are listed below, with the temporal conjunction in boldface, and the modifier underlined.

• adverbial (*just, even, shortly, nearly*):

(97) [But] [even **before** it begins,] [the campaign is drawing fire from anti-smoking advocates,...]$_{wsj\_0326}$

(98) [Interlogic was spun off from Datapoint four years ago,] [shortly **after** Mr. Edelman took control of Datapoint.]$_{wsj\_0333}$

(99) [Thus, optimistic entrepreneurs await a promised land of less red tape] [-- just **as soon as** Uncle Sam gets around to arranging it.]$_{wsj\_1162}$

• temporal noun phrase (*months, a day, two years*):

(100) [Federal officials seized the association in April,] [a day **after** the parent corporation entered bankruptcy-law proceedings.]$_{wsj\_0335}$

(101) [Rep. Gonzalez has complained] [that regulators waited far too long, however,] [ignoring a recommendation from regional officials] [to place Lincoln into receivership] [two years **before** it failed.]$_{wsj\_0335}$

• temporal prepositional phrase (*in* + temporal-NP, *during* + temporal-NP)

(102) [Senator Phil Gramm pointed out Monday] [that] [in the 20 years **before** Gramm-Rudman was enacted in 1985,] [federal spending grew by about 11% a year;] [since the law, it's grown at under 5% annually.]$_{wsj\_1952}$

• another conjunction (*until, since*):

(103) [Stock fund redemptions during the 1987 debacle didn't begin to snowball] [until **after** the market opened on Black Monday.]$_{wsj\_2306}$

• combination of the above types:

(104) [Nearly two months **after** saying] [it had been the victim of widespread fraud,] [MiniScribe Corp. disclosed] [it had a negative net work of $88 million as of July 2...]$_{wsj_0361}$

In stative clauses, a temporal NP (*only nine months*) may not be part of a following temporal clause:

(105) [At the most there is only nine months] [before the LDP fuse burns out.]$_{wsj\_1120}$

A diagnostic to determine whether the expression *only nine months before* is a unit is to reverse the order in the sentence:

(106) *Only nine months before the LDP fuse burns out, at the most there is.

Another stative example:

(107) [But it will be years] [**before** it is clear whether higher rates will offset the payouts for such disasters.]$_{wsj\_1302}$

Temporal phrases are never marked as EDUs, even in cases where the temporal phrase is event-like in its semantic content. This decision was a compromise made to ensure more consistent tagging:

(108) [He said] [he expects U.S. interest rates to decline,] [dragging the dollar down to around 1.80 marks by the end of January **after** a short-lived dash to 1.87 marks by the end of November.]$_{wsj\_0301}$

(109) [Even **before** Mr. Tonkin's broadside, some large dealers said] [they were cutting inventories.]$_{wsj\_0618}$

2.8 Correlative Subordinators

There is a category of subordinating conjunctions, called *correlative subordinators* or *correlatives,* that consist of a combination of two markers, one in the subordinate clause and the other in the superordinate clause. Many, but not all of these, may have a comparative or contrastive function. Examples of correlatives include the following:

more/-er/less...than; as ... as; as much ... as; so ... (that); such ... as; so ... as; such ...(that); no sooner...than; if...then; although...yet/nevertheless; whether...or; the...the

Constructions of this type should be broken into separate EDUs, provided that the subordinate clause contains a verbal element, which can be either a main verb or an auxiliary:

(110) [It was **as** easy] [**as** collecting shells at Malibu.]$_{wsj\_1121}$

(111) [Devotees pass hours,] [watching the lights blink] [and listening to the metal balls ping,] [**as**

much to gamble] [**as** to get a little time to be anonymous, alone with their thoughts.]$_{wsj\_1387}$

(112) [But Mr. Wyss said] [he will watch the numbers] [to get an inkling] [of whether consumers' general buying habits may slack off **as much**] [**as** their auto-buying apparently has.]$_{wsj\_0627}$

(113) [Marni Rice plays the maid with **so much** edge] [**as** to steal her two scenes.]$_{wsj\_1163}$

(114) [Savings and loans reject blacks for mortgage loans twice **as often**] [**as** they reject whites,] [the Office of Thrift Supervision said.]$_{wsj\_1189}$

(115) [Moreover, "it's a lot **cheaper** and **quicker** to buy a plan] [**than** to build one."]$_{wsj\_0317}$

(116) [Adults under age 30 like sports cars, luxury cars, convertibles and imports **far more**] [**than** their elders do.]$_{wsj\_1377}$

(117) [Medieval philosophers used to hold the sensible belief] [that it was **more perfect** to exist] [**than** not to exist,] [and that to exist as a matter of necessity was most perfect of all.]$_{wsj\_1158}$

(118) [There is just **so much** going on] [**that** it's difficult to pick just one factor] [that's driving the market.]$_{wsj\_0671}$

(119) [The problem is **so vast**] [**that** we need to try innovative solutions] [-- in small-scale experiments]$_{wsj\_1107}$

(120) [There is **such** a maze of federal, state and local codes] [**that** "building inspectors are backing away from interpreting them,"] [Mr. Dooling says.]$_{wsj\_1162}$

Comparatives with *enough ... to, too... to* are related to correlatives *so ... that, such ... that,* and likewise should be segmented:

(121) [A private market like this just isn't big **enough**] [**to** absorb all that business.]$_{wsj\_1146}$

(122) So far, the French have failed to win **enough** broad-based support] [**to** prevail].$_{wsj\_2361}$

(123) [There were **too** many phones ringing] [and **too** many things happening] [**to** expect market makers to be as efficient as robots.]$_{wsj\_2379}$

(124) [Friday's stock market sell-off came **too** late for many investors] [**to** act.]$_{wsj\_2306}$

If the second correlative marker occurs in a construction where no verbal element is present (main or auxiliary verb), then it is not broken into a separate EDU. This is consistent with the rule that phrasal comparatives are not separate EDUs:

(125) [It is inhumane to imply that a poor, unemployed woman cannot receive immediate relief for her family at fair prices] [because she does not have **as much** to protect **as** a rich family.]$_{wsj\_1935}$

(126) [He has **too** readily swallowed the case **for** the activist law school culture.]$_{wsj\_1357}$

(127) [But when blacks are getting their loan applications rejected twice **as often as** whites] [-- and in some cities, it is three and four time as often --] [I conclude] [that discrimination is part of the problem.]$_{wsj\_1189}$

(128) [The project has been in and out of the pipeline for **more than** a decade.]$_{wsj\_0606}$

(129) [Meanwhile, the savings-and-loan sector fared **better than** any other Dow Jones industry group.]$_{wsj\_0395}$

2.9 Embedded Discourse Units

An embedded discourse unit is one which has one or both of the following properties:

1) it breaks up a unit which is legitimately an EDU in its own right

2) it modifies a portion of an EDU only, not the entire EDU

When an embedded discourse unit is identified, an embedded relation is used to indicate its relation to the main EDU it modifies. While an embedded discourse unit often occurs in the middle of another EDU, this does not necessarily have to be the case. Relative clauses and other types of nominal post-modifiers, appositives, and parentheticals are examples of embedded constructions that may occur in the middle of a unit, or at the end. We discuss each of these types of embedded units below.

2.9.1 Relative Clauses

Relative clauses are marked as elementary discourse units if they contain a verbal element. This includes full as well as reduced relative clauses:

(130) [A separate inquiry by Chemical cleared Mr. Edelson of allegations] [that he had been lavishly entertained by a New York money broker.]$_{wsj\_0304}$

(131) [Some entrepreneurs say] [the red tape] [they most love to hate] [is red tape] [they would also hate to lose.]$_{wsj\_1162}$

However, do not segment reduced relative clauses that contain an adjective without a verbal element:

(132) [Each $5000 bond carries one warrant, exercisable from Nov. 28, 1989, through Oct. 26, 1994,] [to buy shares at an expected premium of 2 1/2% to the closing share price] [when terms are fixed Oct. 26.]$_{wsj_1161}$

If a relative clause starts with a quantifier, include that as part of the relative clause:

(133) [So far they have issued scores of subpoenas,] [**some of** which went to members of the New York Merc.]$_{wsj\_0664}$

(134) [The Senate bill was stripped of many popular, though revenue-losing, provisions,] [**a number of** which are included in the House-passed bill.]$_{wsj\_2372}$

(135) [In a riveting day of hearings before the House Banking Committee, the examiners described finding shredded documents, a mysterious Panamanian subsidiary, millions of dollars funneled into a Swiss bank, and a complacent attitude by Mr. Wall's deputies,] [**one of** whom was portrayed as acting more like a public-relations man for the thrift than a federal regulator.]$_{wsj\_0335}$

2.9.2 Nominal Postmodifiers with Non-Finite Clause

Nominal postmodifiers containing non-finite verbs are treated as embedded elementary discourse units:

(136) [Mr. Richardson wouldn't offer specifics] [**regarding** Atco's proposed British project,] [but he said] [it would compete for customers with two huge British power generating companies] [that would be formed under the country's plan] [**to privatize** its massive water and electric utilities]$_{wsj\_2309}$

(137) [According to one dealer,] [Japan said] [it has only 40,000 tons of sugar remaining] [**to be shipped** to it this year by Cuba under current commitments.]$_{wsj\_1932}$

(138) ["I have every intention] [**of making** this the best possible show,] [and having it run one hour is the best way to it,"] [said Rod Perth,] [who was named vice president of late night entertainment in August.]$_{wsj\_2395}$

(139) [The results underscore Sears's difficulties] [**in** **implementing** the "everyday low pricing" strategy] [that it adopted in March, as part of a broad attempt] [**to revive** its retailing business.]$_{wsj\_1105}$

2.9.3 Appositives

Appositives are marked as elementary discourse units. The segmentation of an appositive is handled with a SAME-UNIT construction (discussed in Section 2.9.7 below):

(140) [The fact] [**that** this happened two years ago] [and there was a recovery] [gives people some comfort] [**that** this won't be a problem.]$_{wsj\_2345}$

2.9.4 Parentheticals

Parenthetical expressions that occur within the boundaries of a single sentence are generally treated as embedded EDUs, even if they appear at the end of the sentence. (For further discussion on how to handle parentheticals, see Section 2.11.1):

(141) [The Tass news agency said the 1990 budget anticipates income of 429.9 billion rubles] [($US693.4 billion)] [and expenditures of 489.9 billion rubles] [($US790.2 billion).]$_{wsj\_0311}$

If the parenthetical is a separate sentence, or is otherwise independent from the unit it modifies, do not treat the unit as embedded:

(142) ["Most people can't even remember his name."] [(It is John Hart.)]$_{wsj\_1376}$

2.9.5 Other Embedded Units

An embedded comparison may occur if a comparative clause modifies only a portion of a superordinate clause. In the example below, this occurs because *Than You Think* only modifies *The Housing Market Is a Bigger Mess*, but according to our rules, there is no elementary unit boundary between *column* and the following quotation:

(143) [This is in response to George Melloan's Business World column "The Housing Market Is a Bigger Mess] [Than You Think"] [(op-ed page, Sept. 26)]$_{wsj\_1107}$

2.9.6 Coordinate Clauses in Embedded Units

If coordinate clauses appear in embedded units, such as relative clauses or postmodifiers, the clauses are segmented by the usual rules of clausal segmentation:

(144) She signed up,] [starting as an "inside" adjuster,] [who settles minor claims] [and does a lot of work by phone.]$_{file3}$

(145) [The next day,] [as she prepares a $10,000 check for the Hammacks,] [which will cover the cost [of demolishing the house] [and clearing away the debris,] [she jumps at the slightest noise.]$_{file3}$

(146) [At stake was an $80,000 settlement] [involving who should pay what share of clean up costs at the site of a former gas station,] [where underground fuel tanks had leaked] [and contaminated the soil.]$_{file2}$

2.9.7 Same Unit Constructions

An elementary discourse unit may be broken up because of an embedded construction. In the example below, the parenthetical phrase surrounded by dashes is embedded with respect to the flow of the discourse structure. The split EDU is shown in boldface:

(147) [**But maintaining the key components of his strategy**] [-- a stable exchange rate and high levels of imports --] **will consume enormous amounts of foreign exchange.**]$_{wsj\_0300}$

In order to capture this structure, we devised a multinuclear pseudo-relation called SAME-UNIT which serves to indicate that the two discontinuous parts are really one EDU. When an EDU is split by an embedded unit, an embedded relation is used to indicate the relation between the embedded unit and the main unit. The discourse sub-tree for this example is given in Figure 2 below. The parenthetical unit forms the satellite of the relation ELABORATION-ADDITIONAL-E (the "-e" indicates that the relation is embedded). The unit attaches to the first segment of the split EDU, which it modifies.

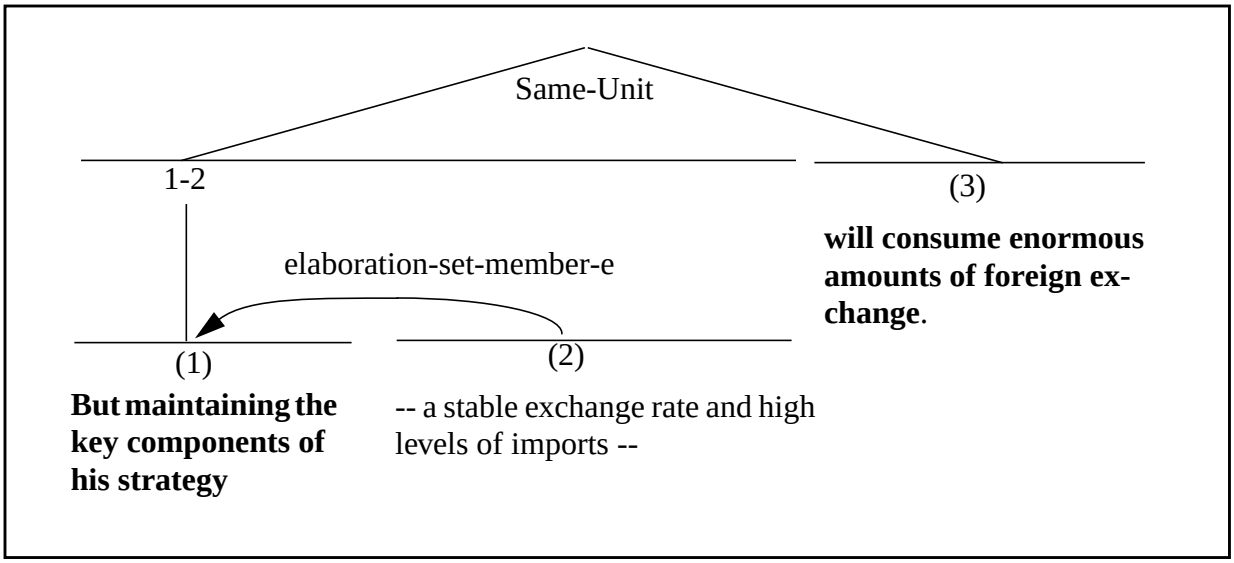

Figure 2: Segmentation of EDUs using SAME-UNIT construction

A different type of embedded unit is illustrated by the next example, in which the clausal modifier *trying to blunt growing demands from Western Europe for a relaxation of controls on exports to the Soviet bloc* is embedded with respect to the EDU *The Bush Administration is questioning*:

> (148) [**The Bush Administration,**] [trying to blunt growing demands from Western Europe for a relaxation of controls on exports to the Soviet bloc,] [**is questioning**] [whether Italy's Ing. C. Olivetti & Co. supplied militarily valuable technology to the Soviets.]$_{wsj\_2326}$

The discourse sub-tree for this example is given in Figure 3. Notice that the embedded clause is attached to the fragment of the split EDU that contains the verb phrase, since it was felt that the relation that best captured the meaning here was PURPOSE-E, a relation which is defined between two events.

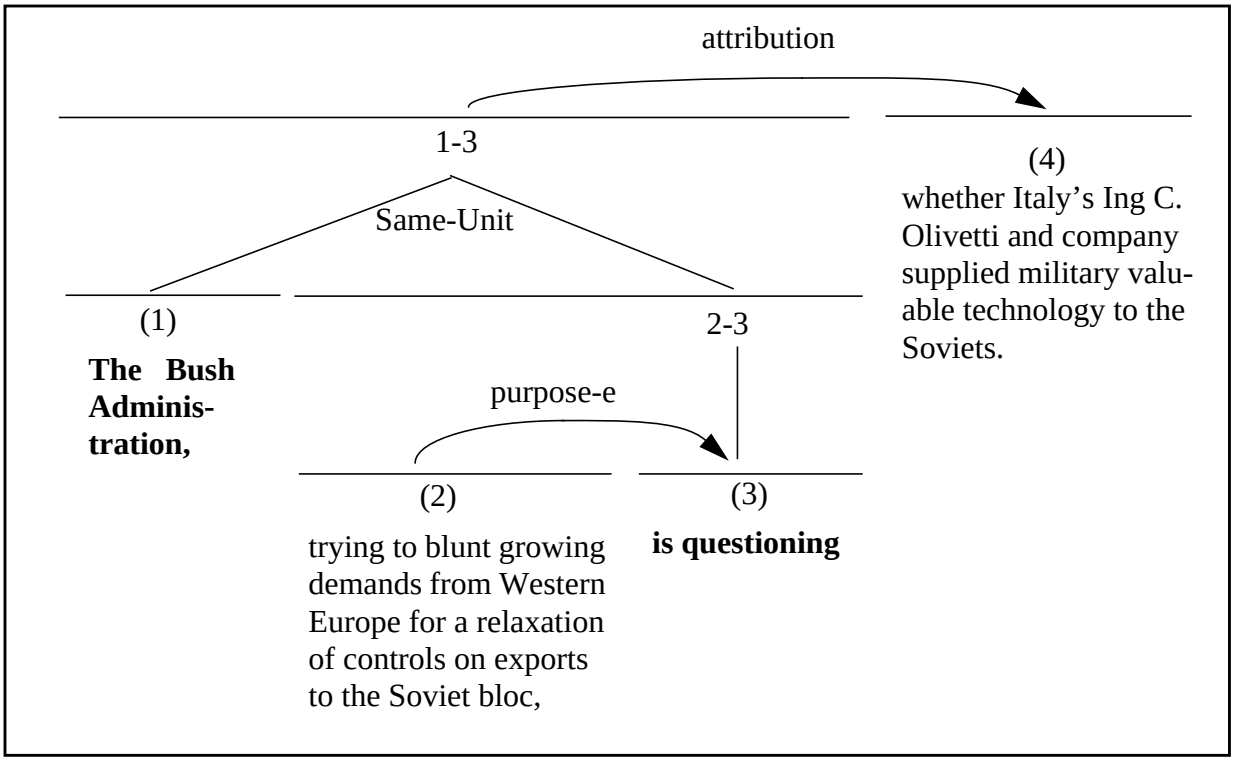

Figure 3: Segmentation of EDUs using SAME-UNIT construction

For verbs of attribution that occur in the middle of a stream of reported speech, select an embedded attribution relation if the source breaks up a single EDU, as in Example (149), but select a non-embedded attribution relation if the source occurs between two separate EDUs, as in (150). In both cases, the attribution satellite is linked to the first segment of the split EDU:

(149) [**"Seeing Michelle up there,"**] [she added,] [**"was like watching myself or my daughter."**]$_{wsj\_0381}$

(150) ["When Sears has a sale at a special price,"] [the woman in the ad declares,] ["it's something] [you don't want to miss."]$_{wsj\_1105}$

The discourse sub-trees for these examples are given in Figures 4 and 5, respectively.

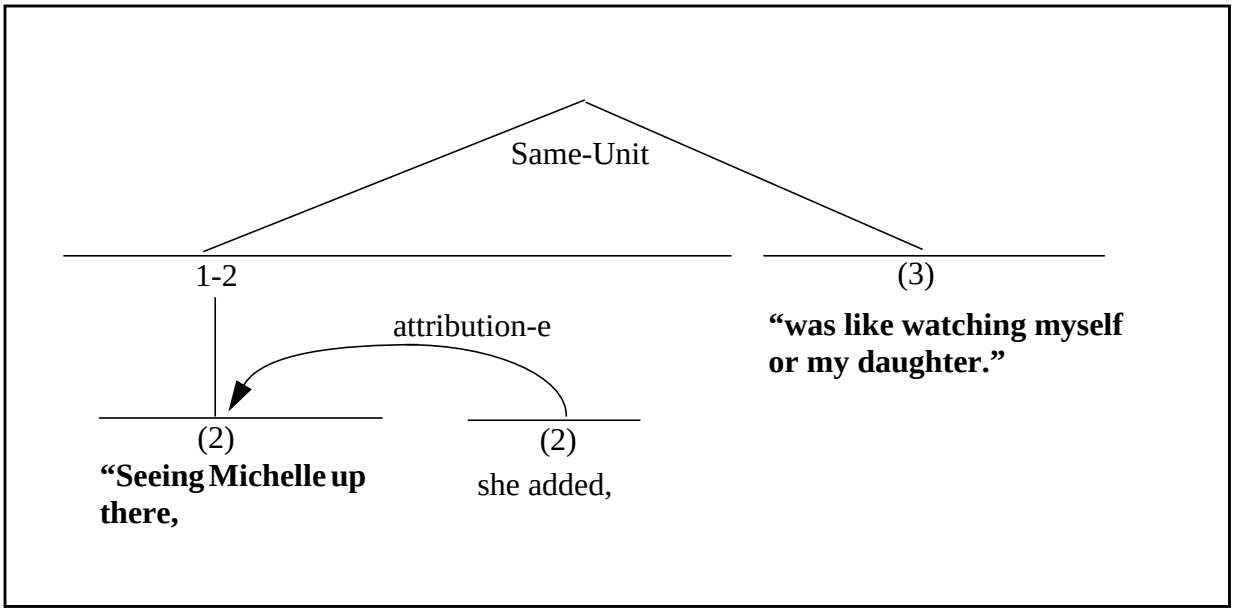

Figure 4: Embedded ATTRIBUTION relation with SAME-UNIT construction.

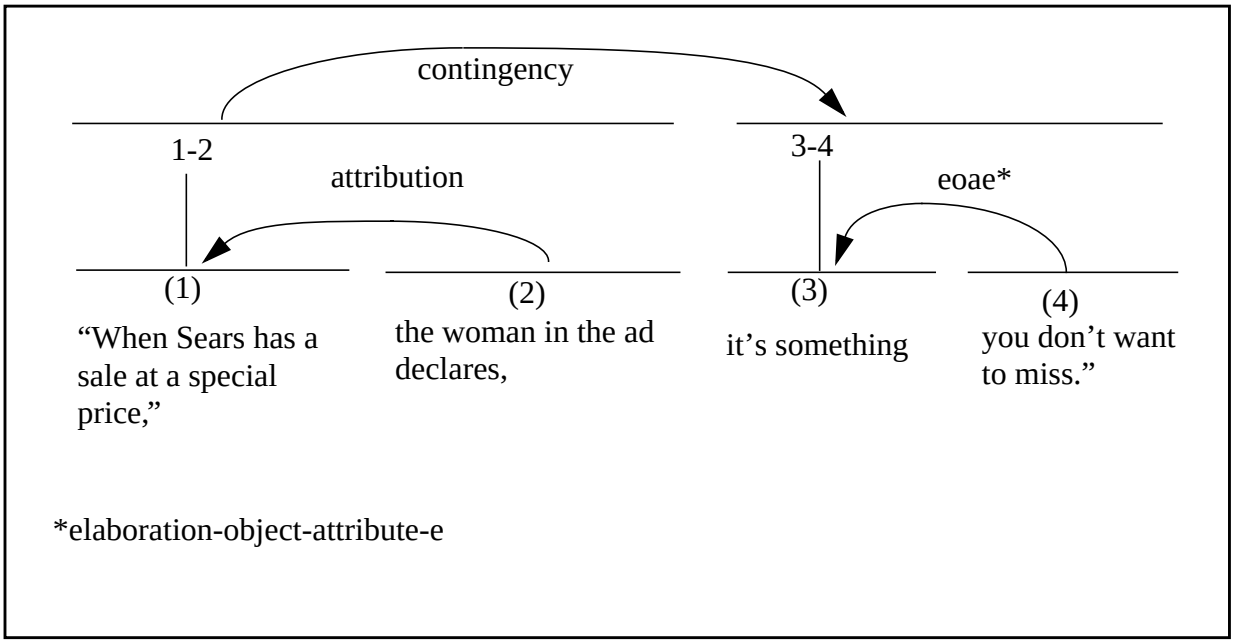

Figure 5: Non-embedded ATTRIBUTION relation with SAME-UNIT construction

2.10 Phrasal Elementary Discourse Units

Phrasal expressions that occur with strong discourse cues are marked as EDUs: We have chosen to enumerate a finite set of discourse cues that justify segmenting a phrase as an EDU: *because, in*

spite of, despite, regardless, irrespective, without, according to, as a result of, not only ... but also. Examples of their usage in the corpus follow:

(151) [Today, no one gets in or out of the restricted area] [**without** De Beers's stingy approval]$_{wsj\_1121}$

(152) [Back then, Mr. Pinter was **not only** the angry young playwright,] [**but also** the first] [to use silence and sentence fragments and menacing stares, almost to the exclusion] [of what we previously understood to be theatrical dialog.]$_{wsj\_1936}$

(153) [**Despite** the yen's weakness with respect to the mark,] [Tokyo traders say] [they don't expect the Bank of Japan to take any action] [to support the Japanese currency on that front.]$_{wsj\_1102}$

(154) [But some big brokerage firms said] [they don't expect major problems] [**as a result of** margin calls.]$_{wsj\_2393}$

Phrasal expressions beginning with prepositions that may have weak or highly ambiguous discourse cues should *not* be tagged as EDUs: *on, in, with, for, within, under, at, during, upon*. Our decision was motivated by the need for highly consistent tagging, and thus we occasionally sacrificed some potentially discourse-relevant phrases in the process.

(155) [Boeing's shares fell $4 Friday] [to close at $57.375 in composite trading on the New York Stock Exchange.]$_{wsj_2308}$

(156) [Canadian Utilities isn't alone in exploring power generation opportunities in Britain, in anticipation of the privatization program.]$_{wsj\_2309}$

(157) [With $500 apiece and an injection of outside capital, they formed Genentech Inc.]$_{wsj_0339}$

Additionally, we decided not to mark as separate EDUs phrases that begin with *in addition to, besides, except, while, if*. The reason for the decision was that in a majority of contexts, we did not find compelling enough reasons to segment these phrases from a discourse perspective:

(158) [And Orangemund boasts attractions besides diamonds.]$_{wsj\_1121}$

(159) "[We're asking the court for a number of things] [he can grant in addition to the thrill of victory,"] [he says.]$_{wsj\_2354}$

(160) [You don’t want to own anything long except gold stocks.]$_{wsj\_0359}$

(161) [That culture,] [carefully crafted by Mr. Smith,] [leaves little, if any, room for unions.]$_{wsj\_1394}$

(162) [But the technology, while reliable, is far slower] [than the widely used hard drives.]$_{wsj\_1971}$

Lastly, phrasal comparatives are not treated as separate EDUs:

(163) [Limiting care won’t be easy or popular.]$_{wsj\_0314}$

(164) [There results compare with net income of $1.8 million, or 44 cents a share, for the corresponding period last year.]$_{wsj_0344}$

2.11 Punctuation

Punctuation can be relevant to determining the discourse structure. A number of conventions have been established for determining EDU boundaries. They are discussed below.

2.11.1 Parentheses

Parenthetical expressions are usually segmented as EDUs, regardless of whether they are a word, phrase, clause or separate sentence. If the parenthetical occurs within the boundaries of a sentence, it is embedded:

(165) [If the government can stick with them,] [it will be able to halve this year’s 120 billion ruble] [**(US$193 billion)**] [deficit.]$_{wsj_0311}$

If the parenthetical is a separate sentence, or is otherwise independent from the unit it modifies, do not treat the unit as embedded:

(166) [“Most people can’t even remember his name.”] [**(It is John Hart.)**]$_{wsj\_1376}$

Occasionally, a parenthetical expression is used to fill in some missing information in the text (although curly brackets are more typically used for this function; see below under 2.11.3). In such a case, the parenthetical should only be segmented if it forms a legitimate EDU within the structure of the sentence. In the example below, the parenthetical is a proper noun, and thus would not be segmented:

(167) [“The interest-rate sensitives aren’t rallying with the rest of the market] [because of fears about what the **(Federal Reserve)** will do,”] [Mr. Jennison said.]$_{wsj\_0327}$

2.11.2 Dashes

When dashes are used to break off parenthetical information, the dashes are included within an embedded EDU, regardless of whether the information occurs in the middle of the sentence or at the end.

(168) [This will require us to define] [-- **and redefine** --] [what is ‘necessary’ or ‘appropriate’ care.]$_{wsj\_0314}$

(169) [Commercial airlines have the lowest accident rate of all transportation modes] [-- **much lower than cars, for example.**]$_{wsj\_0387}$

(170) [I sell] [-- **a little.**]$_{wsj\_1146}$

When dashes are used for a subtitle (often a triple dash), these should be segmented as EDUs, but they should not be treated as embedded. Similarly, an attribution source at the end of a citation should be segmented, but not treated as embedded:

(171) [See: “Centennial Journal:] [100 Years in Business] [**---Look, Up in the Sky, It’s a ...SST, 1969”**] [-- **WSJ, Oct. 5, 1989**]$_{wsj\_2311}$

Occasionally, a dash or double dash may be used in place of a colon, to separate a header from some additional information, a common practice in the Wall Street Journal In that case, the dash is kept with the header:

(172) [**Federal Home Loan Mortgage Corp. --**] [$200 million of stripped mortgage securities] [underwritten by BT Securities Corp.]$_{wsj_2399}$

(173) [**Dow Chemical Co. --**] [$150 million of 8.55% senior notes due Oct. 15, 2009,] [priced at par.]$_{wsj_2399}$

(174) [**CHICAGO -**] [Sears, Roebuck & Co. is struggling] [as it enters the critical Christmas season.]$_{wsj\_1105}$

2.11.3 Brackets

Curly brackets are treated as part of the actual text, because they are used to fill in implied or missing information. Therefore, information enclosed in curly brackets should only be segmented as a separate EDU if it forms a legitimate EDU within the structure of the sentence, as in the first example below, but not the second:

(175) [The studies] [**{on closing the unit}**] [couldn’t be completed until now,”] [he said.]$_{wsj\_2315}$

(176) [Ideally, we’d like to be the operator **{of the project}** and a modest equity investor.]$_{wsj\_2309}$

2.11.4 Commas and Periods

Commas and periods are not independent justification for an EDU boundary. If a unit is a legitimate EDU and it ends with a comma or period, the punctuation is included as part of that EDU:

(177) [The shares represented 66% of his Dun & Bradstreet holdings,] [according to the company.]$_{wsj\_1157}$

Phrases separated by commas are not EDUs:

(178) [Ogden Projects Inc. said] [net income jumped to $6.6 million, or 18 cents a share, in the third quarter.]$_{wsj_0348}$

2.11.5 Colons and Semi-Colons

Text fragments followed by colons are treated as separate EDUs, even when the fragment is a word or phrase, as long as the text that follows the colon provides further elaboration on the topic introduced by the colon:

(179) [**Dollar:**] [142.85 yen, up 0.95; 1.8415 marks, up 0.0075.]$_{wsj\_0350}$

However, in the examples below, the text that follows the colon is a continuation of a single EDU, so a segment boundary is not introduced:

(180) [See: “Centennial Journal:] [100 Years in Business] [**---Look, Up in the Sky, It’s a ...SST, 1969”**] [**-- WSJ, Oct. 5, 1989**]$_{wsj\_2311}$

(181) [Several other reports come before Friday's jobs data, including: the September leading indicators index, new-home sales and October agricultural prices reports due out tomorrow; the October purchasing managers' index and September construction spending and manufacturers' orders on Wednesday; and October chain-store sales on Thursday.]$_{wsj\_0627}$

Phrases separated by semi-colons, like phrases separated by commas, are not EDUs. Clauses separated by semi-colons are EDUs:

(182) [Academy Insurance fell 1/32 to 1 3/16;] [but volume totaled 1.2 million shares.]$_{wsj\_1937}$

2.11.6 Quotation Marks

Quotation marks do not necessarily trigger an EDU boundary. If a sentence quotes a title, for example, this is treated as part of the current EDU:

(183) [This is in response to George Melloan's Business World column "The Housing Market Is a Bigger Mess] [Than You Think"] [(op-ed page, Sept. 26)]$_{wsj\_1107}$

Quotation marks only trigger an EDU if there is a cognitive verb or other attribution satellite (such as *according to...*):

(184) ["In order to restore confidence and ensure the support of our principal lenders,"] [Mr. Bond said,] ["we embarked on fundamental changes in the structure and direction of the group."]$_{wsj\_1195}$

(185) [The issue was oversubscribed] [and "doing very well,"] [according to an official with lead underwriter Morgan Stanley.]$_{wsj\_1322}$

3.0 Determining Nuclearity

In Rhetorical Structure Theory, each unit or textual span that forms part of a relation is characterized by a rhetorical status or nuclearity assignment. A mononuclear relation contains two units (or spans): a *nucleus*, which represents the more salient or essential piece of information in the relation, and a *satellite*, which indicates supporting or background information. A multinuclear relation contains two or more units or spans of equal importance in the discourse, each of which is assigned the role of nucleus. Nuclearity assignment is often determined simultaneously with the assignment of a rhetorical relation. What counts as a nucleus and what counts as a satellite can rarely be determined in isolation. Sometimes, very similar semantic contents can result in different nuclearity assignments, depending on the context, use of cue phrases, etc. For example, in the first example below, both segments are nuclei of a multinuclear relation. However, in the second example, although the semantic content is very similar, the first segment is a satellite and the second one is a nucleus.

(186) [The earnings were fine and above expectations...] [Nevertheless, Salomon's stock fell $1.125 yesterday...]$_{wsj_1124}$

(187) [*Although the earnings were fine and above expectations,*] [Salomon's stock fell $1.125 yesterday.]

In general, nuclear units or spans can be understood by themselves, in isolation of the satellite units that they relate to. Each individual nucleus of a multinuclear relation may not make complete sense in isolation, but it usually makes sense when taken together with the other nuclei siblings. A satellite unit or span often does not make sense by itself. You can, therefore, distinguish between nuclei and satellites by applying the following two tests:

• *Deletion test:* When a satellite of a relation is deleted, the segment that is left, i.e., the nucleus, can still perform the same function in the text, although it may be somewhat weaker. When the nucleus is deleted, the segment that is left is much less coherent.

• *Replacement test:* Unlike the nucleus, a satellite can be replaced with different information without altering the function of the segment.

Embedded units should be always treated as satellites.

4.0 Selecting Rhetorical Relations

Relations are defined to hold between two elementary units if mononuclear, and among two or more elementary units if multinuclear. Relations can be grouped into classes that share some type of rhetorical meaning (such as *cause, elaboration*). Some of these classes contain only one relation, while others contain several relations. Relations are grouped according to these related classes in Section 4.1 below. For the complete inventory of rhetorical relations used in tagging this corpus, see Appendix II.

4.1 Classes of Rhetorical Relations

Once the elementary units of discourse have been determined, adjacent spans are linked together via rhetorical relations creating a hierarchical structure. Relations may be mononuclear or multinuclear. Mononuclear relations hold between two spans and reflect the situation in which one span, the nucleus, is more salient to the discourse structure, while the other span, the satellite, represents supporting information. Multinuclear relations hold among two or more spans of equal weight in the discourse structure. A total of 53 mononuclear and 25 multinuclear relations were used for the tagging of the RST Corpus.

These 78 relations can be partitioned into 16 classes that share some type of rhetorical meaning. These classes and representative members are listed below:

• ***Attribution:*** attribution, attribution-negative

• ***Background:*** background, circumstance

• ***Cause:*** cause, result, consequence

• ***Comparison:*** comparison, preference, analogy, proportion

• ***Condition:*** condition, hypothetical, contingency, otherwise

• ***Contrast:*** contrast, concession, antithesis

• ***Elaboration:*** elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition

• ***Enablement:*** purpose, enablement

• ***Evaluation:*** evaluation, interpretation, conclusion, comment

• ***Explanation:*** evidence, explanation-argumentative, reason

• ***Joint:*** list, disjunction

• ***Manner-Means:*** manner, means

• ***Topic-Comment:*** problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question

• ***Summary:*** summary, restatement

• ***Temporal:*** temporal-before, temporal-after, temporal-same-time, sequence, inverted-sequence

• ***Topic Change:*** topic-shift, topic-drift

In addition, three relations are used to impose structure on the tree: textual-organization, span, and same-unit (used to link parts of units separated by an embedded unit or span).

In some cases, more than one relation may hold between two textual segments. For example, a causal and a temporal relation may hold between two segments simultaneously. Our goal was to label each node in a rhetorical structure tree with only one relation. In previous work by Marcu et al. (1999) a protocol order was established to help the annotator assess what relation to select in ambiguous cases; relations higher on the protocol ranking would be selected before others lower down. In annotating this corpus, we found that analysts preferred a more local comparison of saliency of relations as a criterion. Thus, for example, ELABORATION-ADDITIONAL was regarded as more general, and thus as less salient, than any other elaboration relation (e.g., ELABORATION-SET-MEMBER), TEMPORAL-BEFORE as more general than CONDITION, etc.

4.2 Higher Level Organization of Rhetorical Relations

In this section, we discuss strategies for organizing higher levels of discourse structure. This includes heuristics for grouping relations in parallel structures, and attaching multiple modifiers to a main nucleus.

4.2.1 Elaboration Strategies

Elaboration is one of the most prevalent forms of modification of a nucleus. Often you will find multiple clauses, sentences, or even paragraphs modifying a nucleus. How should these be grouped?

In the text fragment below, the information has been grouped into three sections. Segment A contains the main nucleus (in bold) of the fragment. Segment B elaborates on A, as does Segment C. (Some of the details of the segments have been omitted for clarity.) The analyst must decide how to group this information:

> (188) [Although bullish dollar sentiment has fizzled, **many currency analysts say a massive sell-off probably won't occur in the near future.** *While Wall Street's tough times and lower U.S. interest rates continue to undermine the dollar, weakness in the pound and the yen is expected to offset those factors.* "By default," the dollar probably will be able to hold up pretty well in coming days, says Francoise Soares-Kemp, a foreign-exchange adviser at Credit Suisse. "We're close to the bottom" of the near-term ranges, she contends. ...^{A}]

[With the stock market wobbly and dollar buyers discouraged by signs of U.S. economic weakness and the recent decline in U.S. interest rates that has diminished the attractiveness of dollar-denominated investments, traders say the dollar is still in a precarious position. “They’ll be looking at levels to sell the dollar,” says James Scalfaro, a foreign-exchange marketing representative at Bank of Montreal. While some analysts say the dollar eventually could test support at 1.75 marks and 135 yen, Mr. Scalfaro and others don’t see the currency decisively sliding under support at 1.80 marks and 140 yen soon.^{B}] [Predictions for limited dollar losses are based largely on the pound’s weak state after Mr. Lawson’s resignation and the yen’s inability to strengthen substantially when there are dollar retreats. With the pound and the yen lagging behind other major currencies, “you don’t have a confirmation” that a sharp dollar downturn is in the works, says Mike Malpede, senior currency analyst at Refco Inc. in Chicago. ...^{C}]$_{wsj\_0693}$

There are at least three possible strategies -- 1) have B elaborate on A, and then have C elaborate on AB combined -- this forms a nested elaboration in which A is part of both nuclei in the sub-tree; 2) combine B and C first, and then have BC modify A -- this forms a stratified elaboration in which each elaboration relation has a different nucleus; and 3) treat B and C as independent modifications of A, with no explicit relation between B and C -- this creates two independent or separate modifications to a single nucleus, A. The three modification strategies are illustrated in Figures 6-8 below, respectively.

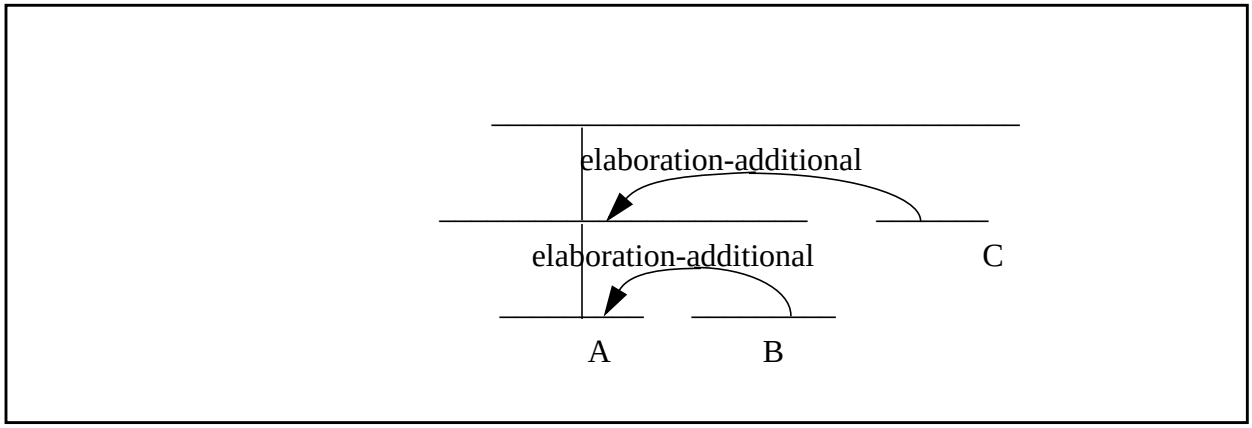

Figure 6: Annotation strategy #1

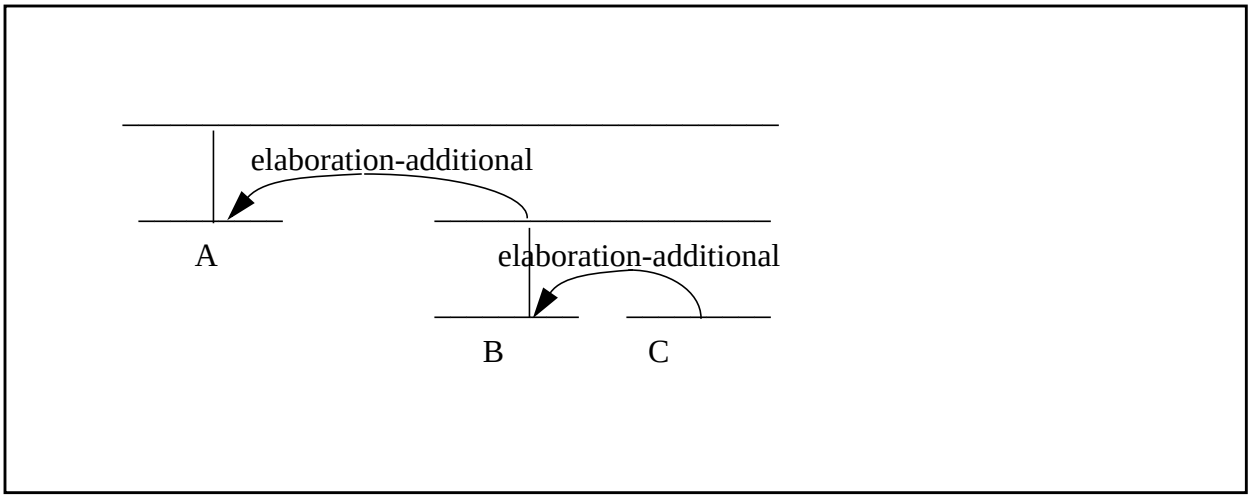

Figure 7: Annotation strategy #2

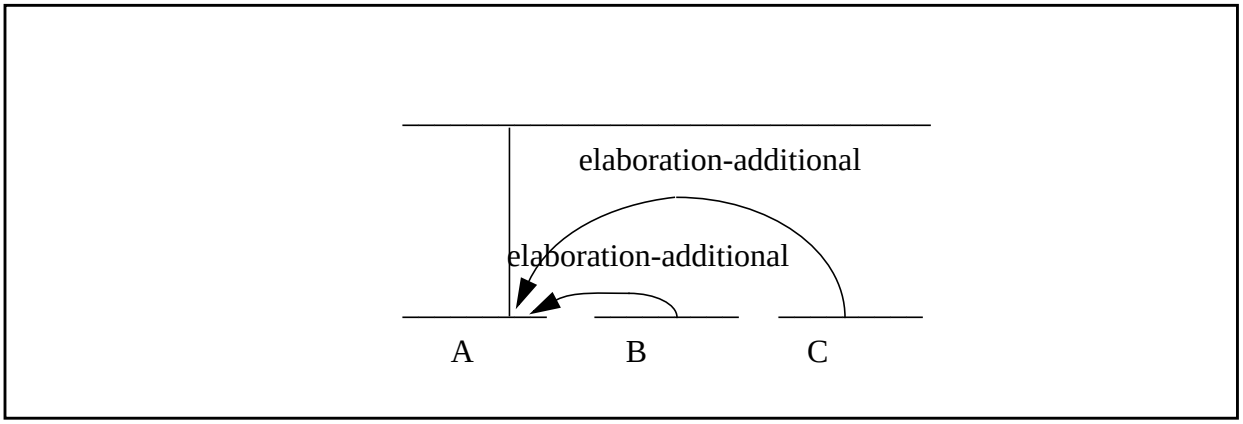

Figure 8: Annotation strategy #3

The annotator selected elaboration strategy #1 for the text fragment in example (188). Segment B elaborates on the “wobbly stock market” and “signs of U.S. economic weakness and the recent decline in U.S. interest rates,” both of which refer back to the portion of Segment A shown in italics in example (188). Segment C starts off by discussing the effect of the “pound’s weak state” and its potential impact on dollar losses. While Segment C also refers back to the portion of Segment A shown in italics, it additionally builds on information presented in Segment B, in which analysts speculate on the possible extent of dollar decline.

If we now take a look at Segment C, and the segments that follow and modify it, namely D and E, we see that a different annotation strategy is appropriate.

(189) [**Predictions for limited dollar losses are based largely on the pound's weak state after Mr. Lawson's resignation and the yen's inability to strengthen substantially when there are dollar retreats.** With the pound and the yen lagging behind other major currencies, "you don't have a confirmation" that a sharp dollar downturn is in the works, says Mike Malpede, senior currency analyst at Refco Inc. in Chicago. ...^{C}] [As far as the pound goes, some traders say a slide toward support at $1.5500 may be a favorable development for the dollar this week. While the pound has attempted to stabilize, currency analysts say it is in critical condition. Sterling plunged about four cents Thursday and hit the week's low of $1.5765 when Mr. Lawson resigned from his six-year post because of a policy squabble with other cabinet members. ... If the pound falls closer to 2.80 marks, the Bank of England may raise Britain's base lending rate by one percentage point to 16%, says Mr. Rendell. But such an increase, he says, could be viewed by the market as "too little too late." The Bank of England indicated its desire to leave its monetary policy unchanged Friday by declining to raise the official 15% discount-borrowing rate that it charges discount houses, analysts say.^{D}] [Pound concerns aside, the lack of strong buying interest in the yen is another boon for the dollar, many traders say. The dollar has a "natural base of support" around 140 yen because the Japanese currency hasn't been purchased heavily in recent weeks, says Ms. Soares-Kemp of Credit Suisse. The yen's softness, she says, apparently stems from Japanese investors' interest in buying dollars against the yen to purchase U.S. bond issues and persistent worries about this year's upheaval in the Japanese government.^{E}]$_{wsj_0693}$

In example (189) Segment C starts out by mentioning two factors for limited dollar losses -- "the pound's weak state" and the "yen's inability to strengthen." Segment D goes on to elaborate exclusively on the situation with the pound, while Segment E elaborates specifically on the situation with the yen. For this example, elaboration strategy #3 is more appropriate, since Segments D and E elaborate independently on C. A test for whether strategy #3 is appropriate is the "deletion test," determining whether the middle segment could be omitted without causing the resulting passage to be incoherent. In this example Segment D could indeed be omitted, and the passage would still be understandable and coherent.

4.2.2 Elaboration with Multinuclear List

Another possible elaboration strategy is when you have a list of items that elaborate on a main point, such as in the following example:

> (190) [The key U.S. and foreign annual interest rates below are a guide to general levels but don't always represent actual transactions.^{A}] [PRIME RATE: 10 1/2%. The base rate on corporate loans at large U.S. money center commercial banks.^{B}] [FEDERAL FUNDS: 8 3/4% high, 8 11/16% low, 8 5/8% near closing bid, 8 11/16% offered. Reserves traded among commercial banks for overnight use in amounts of $1 million or more. Source: Fulton Prebon (U.S.A.) Inc.^{C}] [DISCOUNT RATE: 7%. The charge on loans to depository institutions by the New York Federal Reserve Bank.^{D}] [CALL MONEY: 9 3/4% to 10%. The charge on loans to brokers on stock exchange collateral.^{E}] [COMMERCIAL PAPER placed directly by General Motors Acceptance Corp.: 8.50% 30 to 44 days; 8.25% 45 to 65 days; 8.375% 66 to 89 days; 8% 90 to 119 days; 7.875% 120 to 149 days; 7.75% 150 to 179 days; 7.50% 180 to 270 days.^{F}]$_{wsj_0602}$

In order to capture the structure inherent in the text, the listed items, fragments [B-F] are first linked via the multinuclear LIST relation; then this entire multinuclear structure forms the satellite of an ELABORATION-SET-MEMBER on the text fragment [A] as shown in Figure 9. In general, if modifiers B-F elaborate on A in the same manner, it is reasonable to build a LIST structure first. We typically reserved the LIST relation for cases in which there was some parallel structure across the modifying units, such as parallel syntactic structure, or, as in Example 190, beginning with header information in the same orthographic format. However, in cases where B-F elaborate on A in different ways, or are linked to A with different relations, then using a LIST structure would not be appropriate.

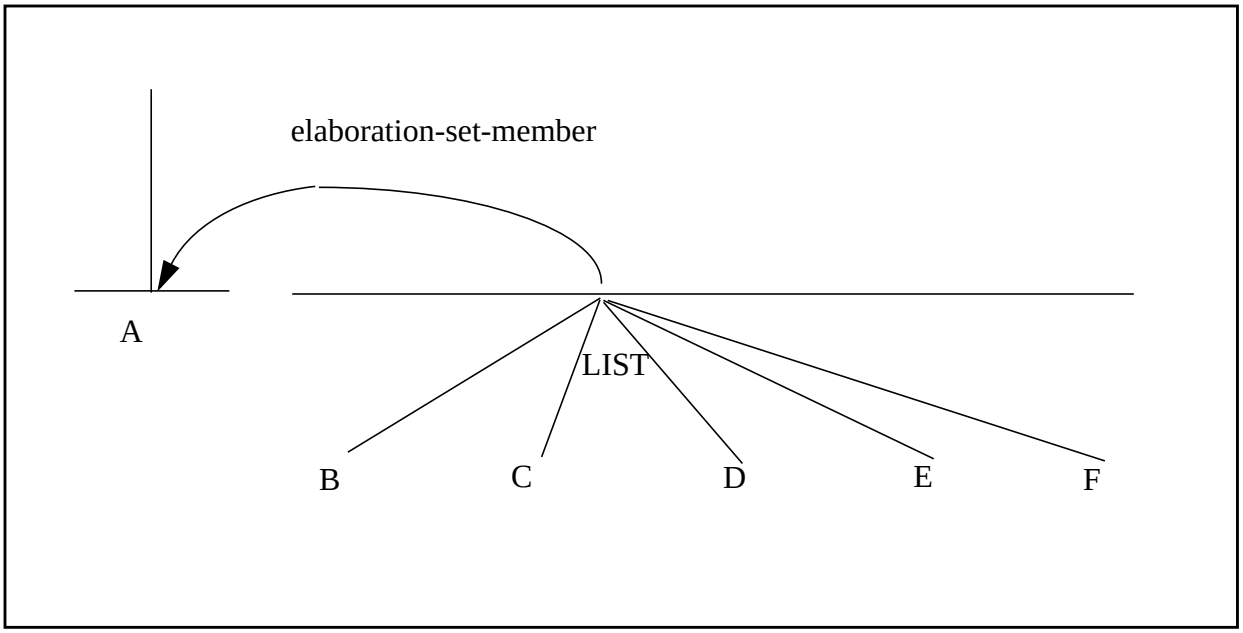

Figure 9: Multinuclear LIST as satellite of ELABORATION-SET-MEMBER relation

Appendix I: Determining EDU Boundaries

The following table presents a list of various syntactic phenomena that have an impact on determining EDU boundaries. The relevant section of the manual is given in parentheses.

Table 1: Syntactic Phenomena and EDUs

Syntactic Unit or Device	EDU?	Qualifications, Exceptions, Examples
Main Clauses (2.1)	Y	Example: *[The company will shut down its plant.]*
Subordinate Clause with Discourse Cue (2.1)	Y	Example: *...[**although** it will not dismiss any employees.]*
Clausal Subjects and Objects (2.2)	N	Example: *[**Shutting down the plant** will be difficult.]*
Clausal Complements (2.3)	N	Exception: complements of attribution verbs **are** EDUs. (see next)
Complements of Attribution Verbs (2.4)	Y	Includes both speech acts and other cognitive acts. • Example: *[The company says] [it will shut down its plant.]* Exception: If the complement is a *to*-infinitival, do not segment. • Example: *[The company wants **to shut down** its plant.]*
Coordinated Sentences (2.5.1)	Y	Example: *[The company will shut down its plant,] [and it will dismiss several hundred employees.]*
Coordination in Superordinate Clauses (2.5.2.1)	Y	Example: *[The company will shut down its plant,] [and dismiss several hundred employees.]*

Syntactic Unit or Device	EDU?	Qualifications, Exceptions, Examples
Coordination in Subordinate Clauses (2.5.2.2)	depends	• If the subordinate construction is normally segmented as an EDU in the single clause case, then the coordinate clauses are segmented as EDUs: • Single clause example: *[The company announced] [that it will shut down its plant].* • Coordinate clause example: *[The company announced] [that it will shut down its plant] [and dismiss several hundred employees.]* If the subordinate construction is not segmented as an EDU in the single clause case, then the coordinate clauses are not segmented as EDUs: • Single clause example: *[The company plans to shut down its plant.]* • Coordinate clause example: *[The company plans to shut down its plant and dismiss several hundred employees.]*
Syntactic Focusing Devices (2.6)	N	When a syntactic focusing device, such as cleft, pseudo-cleft or extraposition creates two clauses out of a single clause, the resulting construction is regarded as a single EDU: • Extraposition example: *[It is hard for the company to dismiss several hundred employees.]*
Temporal Clauses (2.7)	Y	Clausal temporal expressions are EDUs. Temporal clauses triggered by *before, after,* may have a number of modifiers that are included in the EDU: • Example: *[**Just months** **before** dismissing several hundred employees,] ...*
Temporal Phrases (2.7)	N	Temporal phrases, such as *in the morning, in the past several weeks,* are not EDUs. Even if the temporal phrase is event-like in nature, it is not marked as an EDU: • Example: [*Just a week after the company's dismissal of several hundred employees, further layoffs were announced.]*

Syntactic Unit or Device	EDU?	Qualifications, Exceptions, Examples
Correlative Subordinators (2.8)	Y	• Example: *[No sooner had they announced the closing of the plant] [than massive protests erupted on the premises.]*
Embedded Discourse Units (2.9)	Y	Relative clauses, nominal postmodifiers, appositives, parentheticals are treated as embedded EDUs. Embedded units are those which modify a portion of an EDU, or break up another legitimate EDU. • Relative clause example: *[The plant] [that the company will shut down] [is in Ohio.]* • Nominal postmodifier: *[The plant] [(which is in Ohio)] [will be shut down in October.]*
"Discourse-Salient" Phrases (2.10)	Y	Must be marked by a strong discourse cue, such as *according to, because, however.* Phrases marked by cues that are weak or only occasional discourse indicators are not segmented as EDUs: *with, in, besides, during, for.*

Appendix II: Relations Inventory

A total of 53 mononuclear and 25 multinuclear rhetorical relations were used for the tagging of our corpus.[2] In addition, a relation of OTHER (or OTHERMULTINUC) is available as a placeholder, if the annotator is not sure which relation applies in a given circumstance. Table 1 below is a complete listing of all the relations, arranged alphabetically by mononuclear relation. Mononuclear relations are listed in Column 1 if the satellite is the unit that characterizes the relation name. For example, in a BACKGROUND relation, the satellite provides background information for the situation presented in the nucleus. Mononuclear relations listed in Column 2 are those in which the nucleus characterizes the relation name. For example, in a CAUSE relation, the nucleus is the cause of the situation presented in the satellite. Column 3 lists the multinuclear relations. Corresponding mono- and multinuclear relations are shown across a single row. (In some cases, this results in the multinuclear relations appearing out of alphabetical order.)

Table 2: Rhetorical Relations List

Mononuclear (satellite)	Mononuclear (nucleus)	Multinuclear
analogy		Analogy
antithesis		Contrast
attribution		
attribution-n		
background		
	cause	Cause-Result
circumstance		
comparison		Comparison
comment		
		Comment-Topic
concession		
conclusion		Conclusion
condition		

2. The 25 multinuclear relations include two that are not rhetorical relations per se. The SAME-UNIT relation links two non-adjacent parts of a single EDU, e.g., when separated by an intervening relative clause or parenthetical; the TEXTUALORGANIZATION relation links text spans that are marked by schemata labels.

Table 2: Rhetorical Relations List

Mononuclear (satellite)	Mononuclear (nucleus)	Multinuclear
consequence-s	consequence-n	Consequence
contingency		
		Contrast (see antithesis)
definition		
		Disjunction
elaboration-additional		
elaboration-set-member		
elaboration-part-whole		
elaboration-process-step		
elaboration-object-attribute		
elaboration-general-specific		
enablement		
evaluation-s	evaluation-n	Evaluation
evidence		
example		
explanation-argumentative		
hypothetical		
interpretation-s	interpretation-n	Interpretation
		Inverted-Sequence
		List
manner		
means		
otherwise		Otherwise
preference		
problem-solution-s	problem-solution-n	Problem-Solution

Table 2: Rhetorical Relations List

Mononuclear (satellite)	Mononuclear (nucleus)	Multinuclear
		Proportion
purpose		
question-answer-s	question-answer-n	Question-Answer
reason		Reason
restatement		
	result	Cause-Result
rhetorical-question		
		Same-Unit
		Sequence
statement-response-s	statement-response-n	Statement-Response
summary-s	summary-n	
	temporal-before	
temporal-same-time	temporal-same-time	Temporal-Same-Time
	temporal-after	
		TextualOrganization
		Topic-Comment
topic-drift		Topic-Drift
topic-shift		Topic-Shift

Section I.1 below provides an alphabetical listing of all relations used to tag the WSJ corpus, along with their definitions and corresponding examples from the corpus. As a departure from the rest of this manual, where all EDUs in a sentence are bracketed, *only the units relevant to the relation being illustrated are marked by square brackets.* For mononuclear relations, satellites are shown in italics. If additional context is required to help illustrate the example, that will be shown outside of the bracketed EDUs for the relation itself.

II.1. Relations Definitions

Relations are listed alphabetically, followed by a status -- mononuclear, multinuclear, or both. (Note that when a mononuclear and multinuclear relation have the same name, the multinuclear one is distinguished by capitalizing the first letter).

1. ANALOGY (both)

Definition: In an ANALOGY relation, two textual spans, often quite dissimilar, are set in correspondence in some respects. An analogy contains an inference that if two or more things agree with one another in some respects, they will probably agree in other respects. In most cases, the relation is multinuclear.

Mononuclear Example:

(191) *[And just as we did not believe the tendentious claims of the Congressmen and arms-control advocates who visited Krasnoyarsk,]* [we are in no way persuaded by the assent to the tainted-meat theory by a U.S. team of scientists who met with Soviet counterparts in Washington last year.]$_{wsj\_1143}$

2. ANTITHESIS (mononuclear)

Definition: In an ANTITHESIS relation, the situation presented in the nucleus comes in contrast with the situation presented in the satellite. The contrast may happen in only one or few respects, while everything else can remain the same in other respects. An ANTITHESIS relation is always mononuclear -- it is a contrastive relation that distinguishes clearly between the nuclearity of its arguments. It differs from the mononuclear CONCESSION relation, which is characterized by a violated expectation. When both units play a nuclear role, the multinuclear relation CONTRAST should be selected.

Examples:

(192) *[Although the legality of these sales is still an open question,]* [the disclosure couldn't be better timed to support the position of export-control hawks in the Pentagon and the intelligence community.]$_{wsj\_2326}$

(193) *[Tiger itself was founded by a band of gungho airmen who had airlifted supplies over the Hump from India to China during World War II. In the early 1970s, Mr. Smith modeled his fledgling company on Tiger's innovation of hub-and-spoke and containerized-cargo operations.]* [But from early on, Tiger's workers unionized, while Federal's never have.]$_{wsj\_1394}$

3. ATTRIBUTION (mononuclear)

Definition: Instances of reported speech, both direct and indirect, should be marked for the rhetorical relation of ATTRIBUTION. The satellite is the source of the attribution (a clause containing a reporting verb, or a phrase beginning with *according to*), and the nucleus is the content of the reported message (which must be in a separate clause). The ATTRIBUTION relation is also used with cognitive predicates, to include feelings, thoughts, hopes, etc.

Examples:

(194) *[The legendary GM chairman declared]* [that his company would make "a car for every purse and purpose."]$_{wsj\_1377}$

(195) *[Analysts estimated]* [that sales at U.S. stores declined in the quarter, too.]$_{wsj\_1105}$

(196) [The shares represented 66% of his Dun & Bradstreet holdings,] *[according to the company.]*$_{wsj\_1157}$

Counter-examples: In order to segment a sentence into attribution source and content, two conditions must hold:

1) There must be an explicit source of the attribution. If the clause containing the reporting verb does not specify the source of the attribution, and if the source cannot be identified elsewhere in the sentence or nearby context, then a relation of attribution does not hold, and the reporting and reported clauses are treated as one unit. This frequently occurs in passive voice constructions, or generic expressions like *it is said*:

(197) [Earlier yesterday, the Societe de Bourses Francaises was told that a unit of Framatome S.A. also bought Navigation Mixte shares, this purchase covering more than 160,000 share.]$_{wsj\_0340}$

(198) [It is hoped that other Japanese would then follow the leader.]$_{wsj\_0300}$

2) The subordinate clause must not be an infinitival complement. The following examples contain infinitival complements, which are not segmented, and thus, an ATTRIBUTION relation does not hold:

(199) [The former first lady of the Philippines asked a federal court in Manhattan **to dismiss** an indictment against her...]$_{wsj\_0617}$

4. ATTRIBUTION-N (mononuclear)

Definition: This relation marks a negative attribution. The negation must be in the satellite, or source, of the attribution in order for the ATTRIBUTION-N relation to hold. All of the rules for segmenting clauses in the ATTRIBUTION relation also apply to the ATTRIBUTION-N relation.

Examples:

(200) *{Yesterday's statement didn't say]* [whether the Japanese companies will acquire Qintex's remaining stake in the resorts.]$_{wsj\_1372}$

(201) *[I don't know]* [why Barber never told anyone else.]$_{wsj\_1388}$

If the sentence contains an attribution verb which is semantically negative, the ATTRIBUTION-N relation also applies:

(202) *[Bolar has denied]* [that it switched the brand-name product for its own in such testing.]$_{wsj\_2382}$

5. BACKGROUND (mononuclear)

Definition: In a BACKGROUND relation, the satellite establishes the context or the grounds with respect to which the nucleus is to be interpreted. Understanding the satellite helps the reader understand the nucleus. The satellite IS NOT the cause/reason/motivation of the situation presented in the nucleus. The reader/writer intentions are irrelevant in determining whether such a relation holds. In contrast with the CIRCUMSTANCE relation, the information or the context of the BACKGROUND relation is not always specified clearly or delimited sharply. Hence, the CIRCUMSTANCE relation is stronger than BACKGROUND. Often, in a BACKGROUND relation, the events represented in the nucleus and the satellite occur at distinctly different times, whereas events in a CIRCUMSTANCE relation are somewhat co-temporal.

Examples:

(203) *[Banco Exterior was created in 1929 to provide subsidized credits for Spanish exports.]* [The market for export financing was liberalized in the mid-1980s, however, forcing the bank to face competition.]$_{wsj\_0616}$

(204) *[The Voting Rights Act of 1965 was enacted to keep the promise of the Fifteenth Amendment and enable Southern blacks to go to the polls, unhindered by literacy tests and other exclusionary devices.]* [Twenty-five years later, the Voting Rights Act has been transformed by the courts and the Justice Department into a program of racial gerrymandering designed to increase the number of blacks and other minorities -- Hispanics,

Asians and native Americans -- holding elective office.]$_{wsj\_1137}$

6. CAUSE (mononuclear)

Definition: The situation presented in the nucleus is the cause of the situation presented in the satellite. The cause, which is the nucleus, is the most important part. The satellite represents the result of the action. The intention of the writer is to emphasize the cause. When the result is the nucleus, the mononuclear relation RESULT should be selected. When it is not clear whether the cause or result is more important, select the multinuclear relation CAUSE-RESULT.

Example:

(205) [This year, a commission appointed by the mayor to revise New York's system of government completed a new charter,] *[expanding the City Council to 51 from 35 members.]*$_{wsj\_1137}$

7. CAUSE-RESULT (multinuclear)

Definition: This is a causal relation in which two EDUs, one representing the cause and the other representing the result, are of equal importance or weight. When either the cause or the result is more important, select the corresponding mononuclear relation CAUSE or RESULT, respectively.

Example:

(206) To try to combat the traffic slowdown, airlines started reducing fares; average fares rose only 1.7% in August, in contrast to increases of 16% each in February and March. [But so far, the effort has failed,] [and traffic is still slow.]$_{wsj\_1192}$

The multinuclear relations CAUSE-RESULT and CONSEQUENCE are similar. The former should be selected when the causality is perceived as being more direct, while the latter is reserved for a more indirect causal connection.

8. CIRCUMSTANCE (mononuclear)

Definition: In a CIRCUMSTANCE relation, the situation presented in the satellite provides the context in which the situation presented in the nucleus should be interpreted. The satellite IS NOT the cause/reason/motivation of the situation presented in the nucleus. The reader/writer intentions are irrelevant in determining whether such a relation holds. Select CIRCUMSTANCE over BACKGROUND when the events described in the nucleus and satellite are somewhat co-temporal.

Examples:

(207) *[As previously reported,]* [a member of the Philippines' House of Representatives has sued to stop the plant.]$_{wsj\_0606}$

(208) [The project appeared to be on the rocks earlier this month] *[when the other major partner in the project, China General Plastics Corp., backed out. China General Plastics, another Taiwanese petrochemical manufacturer, was to have a 40% stake in Luzon Petrochemical.]*$_{wsj\_0606}$

9. COMMENT (mononuclear)

Definition: In a COMMENT relation, the satellite constitutes a subjective remark on a previous segment of the text. It is not an evaluation or an interpretation. The comment is usually presented from a perspective that is outside of the elements in focus in the nucleus.

Examples:

(209) "There are really very few companies that have adequate capital to buy properties in a raw state for cash. [Typically, developers option property, and then once they get the administrative approvals, they buy it," said Mr. Karatz, adding that he believes the joint venture is the first of its kind.] *["We usually operate in that conservative manner."]*$_{wsj\_2313}$

(210) Sears said [claims from the storm,] *[as expected,]* [reduced its third quarter net by $80 million, or 23 cents a share.]$_{wsj_1105}$

In the example above, the clause *as expected* is a comment, because it represents the perspective of the writer or readers of the article, not the company.

10. COMMENT-TOPIC (multinuclear)

Definition: A specific remark is made on a topic or statement, after which the topic itself is identified. This relation is always multinuclear, as both spans are necessary to understand the context. When the spans occur in the reverse order, with the topic preceding the comment, the relation TOPIC-COMMENT is selected. While COMMENT-TOPIC is not a frequently used device in English, it is seen in news reporting, for example, when someone makes a statement, after which a reference is given to help the reader interpret the context of the statement.

Example:

(211) [We have no independent evidence linking Fatah to any acts of terrorism since Dec. 15, 1988," he said,[1]] [referring to the specific PLO group that Mr. Arafat heads.[2]]$_{wsj\_1101}$

11.COMPARISON (both)

Definition: In a COMPARISON relation, two textual spans are compared along some dimension, which can be abstract. The relations can convey that some abstract entities that pertain to the comparison relation are similar, different, greater-than, less-than, etc. In the case of a comparison relation, the spans, entities, etc. are not in contrast.

Mononuclear Examples:

(212) [Instead of proposing a complete elimination of farm subsidies,] *[as the earlier U.S. proposal did,]* the new package calls for the elimination of only the most tradedistorting ones.$_{wsj\_1135}$

(213) [It said it expects full-year net of 16 billion yen,] *[compared with 15 billion yen in the latest year.]*$_{wsj\_0663}$

Multinuclear Examples:

(214) Few people are aware [that the federal government lends almost as much money] [as it borrows.]$_{wsj\_1131}$

(215) [It was as easy] [as collecting sea shells at Malibu.]$_{wsj\_1121}$

12. CONCESSION (mononuclear)

Definition: The situation indicated in the nucleus is contrary to expectation in the light of the information presented in the satellite. In other words, a CONCESSION relation is always characterized by a violated expectation. (Compare to ANTITHESIS.) In some cases, which text span is the satellite and which is the nucleus do not depend on the semantics of the spans, but rather on the intention of the writer.

Example:

(216) [She and her husband pulled most of their investments out of the market after the 1987 crash,] [*although she still owns some Texas stock.*]$_{wsj\_2386}$

(217) [Still, today's highest-yielding money funds may beat CDs over the next year] [*even if rates fall*]$_{wsj\_0689}$

13. CONCLUSION (both)

Definition: In a CONCLUSION relation, the satellite presents a final statement that wraps up the situation presented in the nucleus. A CONCLUSION satellite is a reasoned judgment, inference,

necessary consequence, or final decision with respect to the situation presented in the nucleus. When the nucleus and satellite are of equal importance, select the multinuclear CONCLUSION.

Mononuclear Example:

(218) China could exhaust its foreign-exchange reserves as early as next year, a Western government report says, unless imports are cut drastically to help narrow the balance-of-payments deficit. According to the report, completed last month, if China's trade gap continues to widen at the pace seen in the first seven months of this year, the reserves would be wiped out either in 1990 or 1991. [A country is considered financially healthy if its reserves cover three months of its imports. The $14 billion of reserves China had in June would cover just that much.] *[The report by the Western government, which declines to be identified, concludes that "a near-term foreign-exchange payment problem can be avoided only if import growth drops to below 5% per annum."]*wsj_1391

14. CONDITION (mononuclear)

Definition: In a CONDITION relation, the truth of the proposition associated with the nucleus is a consequence of the fulfillment of the condition in the satellite. The satellite presents a situation that is not realized.

Examples:

(219) [S.A. brewing would make a takeover offer for all of Bell Resources] *[if it exercises the option,]* according to the commission.wsj_0630

Negative conditions may also be represented with this relation:

(220) A company spokesman said [the gain on the sale couldn't be estimated] *[until the "tax treatment has been determined."]*wsj_1179

(221) [However, competitors say that Kidder's hiring binge involving executive-level staffers, some with multiple-year contract guarantees, could backfire] *[unless there are results.]*wsj_0604

15. CONSEQUENCE (multinuclear), CONSEQUENCE-N (mononuclear), CONSEQUENCE-S

(mononuclear)

Definition: In a consequence relation, the situation presented in one span is a consequence of the situation presented in the other span. The reader/writer intentions are irrelevant to determining whether such a relation holds. A CONSEQUENCE-N relation is similar to a RESULT relation, in that in both cases, the *nucleus* presents a consequence or result of the situation in the satellite. Similarly, a CONSEQUENCE-S relation is similar to a CAUSE relation, in that in both cases, the *satellite* presents a consequence or result of the situation in the nucleus. The relations CAUSE and RESULT imply a more direct linkage between the events in the nucleus and the satellite, whereas a CONSEQUENCE-S or CONSEQUENCE-N relation suggests a more indirect linkage. If both spans carry equal weight in the discourse, select the multinuclear CONSEQUENCE.

Example of CONSEQUENCE-N:

(222) *[There is such a maze of federal, state and local codes]* [that "building inspectors are backing away from interpreting them,"] Mr. Dooling says.$_{wsj\_1162}$

Examples of CONSEQUENCE-S:

(223) [This hasn't been Kellogg Co.'s year. The oat-bran craze has cost the world's largest cereal maker market share.] *[The company's president quit suddenly.]*$_{wsj\_0610}$

(224) The Dow Jones Industrial Average tumbled more than 60 points after the report's release, [before recovering] *[to close 18.65 points lower at 2638.73.]*$_{wsj\_1970}$

Example of CONSEQUENCE:

(225) [Many a piglet won't be born as a result,] [and many a ham will never hang in a butcher shop.]$_{wsj\_1146}$

16. CONTINGENCY (mononuclear)

Definition: In a CONTINGENCY relation, the satellite suggests an abstract notion of recurrence or habituality. Hence, the expression of time, place, or condition is not the primary focus.

Examples:

(226) [Today, no one gets in or out of the restricted area] *[without De Beers's stingy approval.]*$_{wsj\_1121}$

(227) [They have a life of their own and can be counted on to look good and perform] *[whenever a cast isn't up to either.]*$_{wsj\_1154}$

17. CONTRAST (multinuclear)

Definition: In a CONTRAST relation, two or more nuclei come in contrast with each other along some dimension. The contrast may happen in only one or few respects, while everything else can remain the same in other respects. Typically, a CONTRAST relation includes a contrastive discourse cue, such as *but, however, while,* whereas a COMPARISON does not.

Examples:

(228) [But from early on, Tiger's workers unionized,] [while Federal's never have.]$_{wsj\_1394}$

(229) [The proposal reiterates the U.S. desire to scrap or reduce a host of trade-distorting subsidies on farm products.] [But it would allow considerable flexibility in determining how and when these goals would be achieved.]$_{wsj\_1135}$

18. DEFINITION (mononuclear)

Definition: In a DEFINITION relation, the satellite gives a definition of the nucleus.

Example:

(230) [Deciding what constitutes "terrorism" can be a legalistic exercise.] *[The U.S. defines it as "premeditated, politically motivated violence perpetrated against noncombatant targets by subnational groups or clandestine state agents."]*$_{wsj\_1101}$

19. DISJUNCTION (multinuclear)

Definition: DISJUNCTION is a multinuclear relation whose elements can be listed as alternatives, either positive or negative.

Examples:

(231) [Call it a fad.] [Or call it the wave of the future.]$_{wsj\_0633}$

(232) [Yet Israel will neither share power with all these Arabs] [nor, says its present prime minister, redraw its borders closer to its pre-1967 Jewish heartland.]$_{wsj\_1141}$

20. ELABORATION-ADDITIONAL (mononuclear)

Definition: In an ELABORATION-ADDITIONAL relation, the satellite gives additional information or detail about the situation presented in the nucleus. This relation is extremely common at all levels of the discourse structure, and is especially popular to show relations across large spans

of information. It is the default for the family of elaboration relations, and should be used when none of the other, more specific, elaboration relations apply.

Examples:

(233) [UNDER A PROPOSAL by Democrats to expand Individual Retirement Accounts, a $2,000 contribution by a taxpayer in the 33% bracket would save $330 on his taxes.] *[The savings was given incorrectly in Friday's edition.]*$_{wsj_0605}$

(234) [The company wouldn't elaborate,] *[citing competitive reasons.]*$_{wsj\_0610}$

When an elaborating satellite is an embedded unit, the relation ELABORATION-ADDITIONAL-E is used. Generally, the embedded satellite of this relation is separated from the nucleus by a comma and represents a nonrestrictive relative clause, whereas an ELABORATION-OBJECT-ATTRIBUTE-E represents a restrictive relative clause:

(235) [Prices for the machine,] *[which can come in various configurations,]* [are $2 million to $10 million.]$_{wsj_2396}$

21. ELABORATION-GENERAL-SPECIFIC (mononuclear)

Definition: The satellite provides specific information to help define a very general concept introduced in the nucleus.

Examples:

(236) [The projects are big.] *[They can be C$1 billion plus.]*$_{wsj_2309}$

(237) Proponents maintain that a president would choose to use a line-item veto more judiciously than that. [But there may be another problem with the device:] *[Despite all the political angst it would cause, it mightn't be effective in cutting the deficit. Big chunks of the government budget, like the entitlement programs of Social Security and Medicare, wouldn't be affected.]*$_{wsj\_0609}$

22. ELABORATION-OBJECT-ATTRIBUTE (mononuclear)

Definition: ELABORATION-OBJECT-ATTRIBUTE is a relation involving a clause, usually a postmodifier of a noun phrase, that is required to give meaning to an animate or inanimate object. The modifying clause is the satellite, and the object it modifies is the nucleus. The satellite is *intrinsic* to the meaning of the nucleus, and/or essential to understanding the context, and is almost always an embedded unit. This relation is almost always embedded.

In embedded units, ELABORATION-OBJECT-ATTRIBUTE-E is much more common than ELABORATION-ADDITIONAL-E. The way to distinguish between ELABORATION-OBJECT-ATTRIBUTE-E and ELABORATION-ADDITIONAL-E is that the former is intrinsic to the object it is modifying in the nucleus, while the latter is extrinsic, i.e. it contains information that is incidental and which could be omitted without loss of understanding. Usually, an ELABORATION-ADDITIONAL-E is separated from the nucleus by a comma and represents a nonrestrictive relative clause, whereas an ELABORATION-OBJECT-ATTRIBUTE-E represents a restrictive relative clause.

Examples:

(238) [Allied Capital is a closed-end management investment company] [*that will operate as a business development concern.*]$_{wsj\_0607}$

The relative clause further characterizes the type of “closed-end management investment company” being discussed, and is therefore essential information in the context.

(239) [The magnitude of the exchange's problems may not become known for some time^{1}] [because of Lloyd's practice^{2}] *[of leaving the books open for three years to allow for the settlement of the claims.*^{3}*]*$_{wsj\_1302}$

Leaving the books open for three years *is* the practice, and therefore this relation is ELABORATION-OBJECT-ATTRIBUTE-E, not ELABORATION-ADDITIONAL-E. The relation is embedded because the satellite, unit [3], modifiers only the portion of unit [2] underlined.

(240) [That year, it posted record pretax profit of $650 million, a gain] *[it attributes to higher rates and fewer claims.]*$_{wsj_1302}$

In this example, *gain* is a very generic, abstract noun. The modifying clause is necessary to understanding the context, so the relation is ELABORATION-OBJECT-ATTRIBUTE-E.

23. ELABORATION-PART-WHOLE (mononuclear)

Definition: In an ELABORATION-PART-WHOLE relation, the satellite specifies or elaborates on a portion or part of the nucleus. Since this relation is most often appropriate for representing parts of objects, it is only occasionally found to hold between EDUs. Most typically, this would be in a parenthetical modifier. It is distinguished from ELABORATION-SET-MEMBER, in which each member is representative of the entire set in a similar way.

Example:

(241) [The pride of Orangemund is the 18-hole golf course] *[-- with the largest sand trap in the world.]*$_{wsj\_1121}$

24. ELABORATION-PROCESS-STEP (mononuclear)

Definition: In this elaboration relation, the nucleus introduces an activity or event (a process). The satellite then enumerates the steps involved in carrying out the process, usually in chronological order. Thus, the steps are most often represented in a multinuclear SEQUENCE relationship.

Example:

(242) [Montedison S.p. A. definitively agreed to buy all of the publicly held shares of Erbamont N.V. for $37 each. Montedison now owns about 72% of Erbamont's shares outstanding. The companies said the accord was unanimously approved by a special committee of Erbamont directors unaffiliated with Montedison.[1]] *[Under the pact, Montedision will make a $37-a-share tender offer for Erbamont stock outstanding.[2]] [The tender offer will be followed by the sale of all of Erbamont's assets, subject to all of its liabilities, to Montedison.[3]] [Erbamont will then be liquidated, with any remaining Erbamont holders receiving a distribution of $37 a share.[4]]*$_{wsj_0660}$

In the above example, spans [2] and [3] and [4] are the steps, which are combined in a multinuclear SEQUENCE relation to form the satellite of an ELABORATION-PROCESS-STEP relationship for the situation mentioned in span [1], the nucleus.

25. ELABORATION-SET-MEMBER (mononuclear)

Definition: In this elaboration relation, the nucleus introduces a finite set (which may be generic or a named entity) or a list of information. The satellite then specifically elaborates on at least one member of the set. Typically, the members themselves are represented in a multinuclear LIST relationship.

Examples:

(243) [The individuals said gulf power and federal prosecutors are considering a settlement under which the company would plead guilty to two felony charges] and pay fines totaling between $500,000 and $1.5 million.[1]] *[Under one count, Gulf power would plead guilty to conspiring to violate the Utility Holding Company Act.[2]] [Under the second count, the company would plead guilty to conspiring to evade taxes.[3]]*$_{wsj_0619}$

In the above example, spans [2] and [3] are the set members, which are combined in a LIST relation to form the satellite of an ELABORATION-SET-MEMBER relationship for the felony charges mentioned in span [1], the nucleus.

(244) [National also participates in the Northwest frequent-flyer program along with four other airlines,[1]] [including Delta and USAir Group Inc.'s USair unit.[2]]$_{wsj\_2394}$

Unit [2] is an ELABORATION-SET-MEMBER-E of the phrase *four other airlines* in the nucleus.

26. ENABLEMENT (mononuclear)

Definition: In an ENABLEMENT relation, the situation presented in the nucleus is unrealized. The action presented in the satellite increases the chances of the situation in the nucleus being realized.

Examples:

(245) ["I'd like to see (Kidder) succeed.] *[But they have to attract good senior bankers who can bring in the business from day one."]*$_{wsj\_0604}$

(246) *[The administration of federal credit should closely parallel private lending practices, including the development of a loan loss reserve and regular outside audits.]* [Establishing these practices would permit earlier identification of emerging financial crises, provide better information for loan sales and budgeting decisions, and reduce fraud.]$_{wsj\_1131}$

27. EVALUATION (multinuclear), EVALUATION-N (mononuclear), EVALUATION-S (mononuclear)

Definition: In an evaluation relationship, one span assesses the situation presented in the other span of the relationship on a scale of good to bad. An evaluation can be an appraisal, estimation, rating, interpretation, or assessment of a situation. The evaluation can be the viewpoint of the writer or another agent in the text. The assessment may occur in the satellite (EVALUATION-S) or the nucleus (EVALUATION-N), or it may occur in a multinuclear relationship (EVALUATION), when the spans representing the situation and the assessment are of equal weight.

Example of EVALUATION-N:

(247) *[What defeated General Aoun was not only the weight of the Syrian army. The weight of Lebanon's history was also against him;]* [and it is a history Israel is in danger of repeating.]$_{wsj\_1141}$

Example of EVALUATION-S:

(248) [But racial gerrymandering is not the best way to accomplish that essential goal.] *[It is a quick fix for a complex problem.]*$_{wsj\_1137}$

Example of EVALUATION:

(249) [Employers must deposit withholding taxes exceeding $3,000 within three days after payroll -- or pay stiff penalties --] [and that's a big problem for small businesses.]$_{wsj_1162}$

28. EVIDENCE (mononuclear)

Definition: In an EVIDENCE relation, the situation presented in the satellite provides evidence or justification for the situation presented in the nucleus. Usually EVIDENCE relations pertain to actions and situations that are independent of the will of an animate agent. Evidence is data on which judgment of a conclusion may be based, and is presented by the writer or an agent in the article to convince the reader of a point. An evidence satellite increases the chance of the reader accepting the information presented in the nucleus.

Examples:

(250) [That system has worked.] *[The standard of living has increased steadily over the past 40 years; more than 90% of the people consider themselves middle class.]*$_{wsj\_1120}$

(251) [The gap between winners and laggards will grow.] *[In personal computers, Apple, Compaq and IBM are expected to tighten their hold on their business. At the same time, second-tier firms will continue to lose ground. Some lagging competitors even may leave the personal computer business altogether. Wyse Technology, for instance, is considered a candidate to sell its troubled operation. "Wyse has done well establishing a distribution business, but they haven't delivered products that sell," said Kimball Brown, an analyst at Prudential-Bache Securities. Mr. Brown estimates Wyse, whose terminals business is strong, will report a loss of 12 cents a share for its quarter ended Sept.]*$_{wsj\_2365}$

29. EXAMPLE (mononuclear)

Definition: The satellite provides an example with respect to the information presented in the nucleus. If the example is a member of a set which may be enumerable but not all of whose elements are known or specified, then choose the relation EXAMPLE. If the example is a member of

an enumerable set whose elements are known or specified, select the relation ELABORATION-SET-MEMBER.

Examples:

(252) [The fiscal 1990 measure builds on a pattern set earlier this year by House and Senate defense authorizing committees, and -- at a time of retrenchment for the military and concern about the U.S.'s standing in the world economy -- overseas spending is most vulnerable.] *[Total Pentagon requests for installations in West Germany, Japan, South Korea, the United Kingdom and the Philippines, for example, are cut by almost two-thirds, while lawmakers added to the military budget for construction in all but a dozen states at home.]*$_{wsj\_0686}$

(253) [The offer is based on several conditions,] [including obtaining financing.]$_{wsj\_0661}$

30. EXPLANATION-ARGUMENTATIVE (mononuclear)

Definition: An EXPLANATION-ARGUMENTATIVE relation usually pertain to actions and situations that are independent of the will of an animate agent.The satellite provides a *factual* explanation for the situation presented in the nucleus. It is not the intention of the writer to convince the reader of a point, which is the role of the EVIDENCE relation. It also differs from the REASON relation, which justifies or explains the actions of an animate agent, and involves the will or intentions of the agent.

Example:

(254) [But their 1987 performance indicates that they won't abandon stocks unless conditions get far worse.] *[“Last time, we got rewarded for going out and buying stocks when the panic was the worst,” said John W. Rogers, president of Chicago-based Ariel Capital Management Inc., which manages $1.1 billion of stocks.”]*$_{wsj_2381}$

31. HYPOTHETICAL (mononuclear)

Definition: In a HYPOTHETICAL relation, the satellite presents a situation that is not factual, but that one supposes or conjectures to be true. The nucleus presents the consequences that would arise should the situation come true. A HYPOTHETICAL relation presents a more abstract scenario than a CONDITION relation. This relation is always mononuclear.

Example:

(255) *[Theoretically, the brokers will then be able to funnel “leads” on corporate finance opportunities to*

Kidder's investment bankers,] [possibly easing the longstanding tension between the two camps.]$_{wsj\_0604}$

32. INTERPRETATION (multinuclear), INTERPRETATION-N (mononuclear), INTERPRETATION-S (mononuclear)

Definition: In interpretation relations, one side of the relation gives a different perspective on the situation presented in the other side. It is subjective, presenting the personal opinion of the writer or of a third party. An interpretation can be: 1) an explanation of what is not immediately plain or explicit; 2) an explanation of actions, events, or statements by pointing out or suggesting inner relationships, motives, or by relating particulars to general principles; or 3) an understanding or appreciation of a situation in light of individual belief, judgment, interest, or circumstance.

The interpretation may be mononuclear, with the interpretation occurring in the satellite (INTERPRETATION-S) or in the nucleus (INTERPRETATION-N); or it may be multinuclear (INTERPRETATION), with the interpretation occurring in one of the nuclei.

Example of INTERPRETATION-N:

(256) *[Even while they move outside their traditional tony circle, racehorse owners still try to capitalize on the elan of the sport. Glossy brochures circulated at racetracks gush about the limelight of the winner's circle and high-society schmoozing. One handout promises: "Pedigrees, parties, post times, parimutuels and pageantry."]* ["It's just a matter of marketing and promoting ourselves," says Headley Bell, a fifth-generation horse breeder from Lexington.]$_{wsj\_1174}$

Examples of INTERPRETATION-S:

(257) [Far from promoting a commonality of interests among black, white, Hispanic and other minority voters, drawing the district lines according to race suggests that race is the voter's and the candidate's most important trait.] *[Such a policy implies that only a black politician can speak for a black person, and that only a white politician can govern on behalf of a white one.]*$_{wsj\_1137}$

The next example illustrates two embedded relations, both of which are INTERPRETATION-S-E:

(258) [Foreclosed homes could be sold by the FHA for no down payment] *[(the biggest obstacle to young buyers),]* [but with personal liability for the mortgage] *[(no walking away by choice).]* $_{wsj\_1107}$

Example of INTERPRETATION:

(259) But John LaWare, a Fed governor, told the sub-committee [the evidence is mixed] [and that the Fed's believes (sic) the vast majority of banks aren't discriminating.]$_{wsj\_1189}$

33. INVERTED-SEQUENCE (multinuclear)

Definition: An INVERTED-SEQUENCE is a multinuclear list of events presented in reverse chronological order. (See also SEQUENCE).

Example:

(260) [Three new issues begin trading on the New York Stock Exchange today,] [and one began trading on the Nasdaq/National Market System last week.]$_{wsj\_0607}$

34. LIST (multinuclear)

Definition: A LIST is a multinuclear relation whose elements can be listed, but which are not in a comparison, contrast or other, stronger type of multinuclear relation. A LIST relation usually exhibits some sort of parallel structure between the units involved in the relation. At lower levels of the discourse structure, such as between clauses or sentences, a LIST relation is often selected when there is some sort of parallel syntactic or semantic structure between the units, such as in the examples below. At higher levels of the discourse structure, the relation may be found when there are paragraphs of items enumerated in a similar fashion (see discussion in Section 4.2.2 above).

Examples:

(261) [A union, sooner or later, has to have an adversary,] [and it has to have a victory.]$_{wsj\_1394}$

(262) [The election, which would bring the first major union to Federal's U.S. operations, has pitted new hires against devoted veterans such as Mr. Brown.] [It has also rattled Federal's strongly anti-union management, which is already contending with melding far-flung operations and with falling profits.]$_{wsj\_1394}$

35. MANNER (mononuclear)

Definition: A manner satellite explains the way in which something is done. (It sometimes also expresses some sort of similarity/comparison.) The satellite answers the question “in what manner?” or “in what way?”. A MANNER relation is less “goal-oriented” than a MEANS relation, and often is more of a description of the style of an action.

Examples:

(263) [Magazine editors did not take the criticisms] [lying down.]$_{wsj\_1123}$

(264) Soon after the merger, moreover, Federal's management asked Tiger's pilots to sign an agreement stating [that they could be fired any time,] [without cause or notice.]$_{wsj\_1394}$

The following example illustrates an embedded, MANNER-E relation:

(265) [In this week's show, there's an unsafe nuclear weaponsmaking facility] [(a la Rocky Flats).]$_{wsj\_1397}$

36. MEANS (mononuclear)

Definition: A means satellite specifies a method, mechanism, instrument, channel or conduit for accomplishing some goal. It should tell you how something was or is to be accomplished. In other words, the satellite answers a "by which means?"or "how?" question that can be assigned to the nucleus. It is often indicated by the preposition *by.*

Example:

(266) [Some underwriters have been pressing for years to tap the low-margin business] *[by selling some policies directly to consumers.]*$_{wsj\_1302}$

37. OTHERWISE (both)

Definition: This is a mutually exclusive relation between two elements of equal importance. The situations presented by both the satellite and the nucleus are unrealized. Realizing the situation associated with the nucleus will prevent the realization of the consequences associated with the satellite. This relation may also be multinuclear.

Mononuclear Example:

(267) [The executive close to Saatchi & Saatchi said that "if a bidder came up with a ludicrously high offer, a crazy offer which Saatchi knew it couldn't beat, it would have no choice but to recommend it to shareholders.] [*But {otherwise} it would undoubtedly come back" with an offer by management.*]$_{wsj\_2331}$

38. PREFERENCE (mononuclear)

Definition: The relation compares two situations, acts, events, etc., and assigns a clear preference for one of the situations, acts, events, etc. The preferred situation, act, event, etc. is the nucleus.

Example:

(268) *[She has thrown extravagant soirees for crowds of people,]* [but prefers more intimate gatherings.]$_{wsj\_1367}$

39. PROBLEM-SOLUTION (multinuclear), PROBLEM-SOLUTION-N (mononuclear), PROBLEM-SOLUTION-S (mononuclear)

Definition: In a problem-solution relation, one textual span presents a problem, and the other text span presents a solution. The relation may be mononuclear or multinuclear, depending on the context. When the problem is perceived as more important than the solution, the problem is assigned the role of nucleus and the solution is the satellite. The relation PROBLEM-SOLUTION-S should be selected in this case. When the solution is the nucleus, use the label PROBLEM-SOLUTION-N; when the relation is multinuclear, use the relation PROBLEM-SOLUTION.

Mononuclear (PROBLEM-SOLUTION-S) Example:

(269) [Despite the drop in prices for thoroughbreds, owning one still isn't cheap. At the low end, investors can spend $15,000 or more to own a racehorse in partnership with others. At a yearling sale, a buyer can go solo and get a horse for a few thousand dollars. But that means paying the horse's maintenance; on average, it costs $25,000 a year to raise a horse.] [*For those looking for something between a minority stake and total ownership, the owners' group is considering a special sale where established horse breeders would sell a 50% stake in horses to newcomers*.]$_{wsj_1174}$

Mononuclear (PROBLEM-SOLUTION-N) Example:

(270) [*But while they may want to be on the alert for similar buying opportunities now, they're afraid of being hammered by another terrifying plunge.*] [The solution, at least for some investors, may be a hedging technique that's well-known to players in the stock-option market. Called a married put, the technique is carried out by purchasing a stock and simultaneously buying a put option on that stock.]$_{wsj\_1962}$

Multinuclear Example:

(271) [However, about 120 employees will be affected by the agreement.] [First Tennessee, assisted by IBM, said it will attempt to place the employees within the company, IBM or other companies in Memphis. The process will take as many as six months to complete, the company said.]$_{wsj\_0621}$

40. PROPORTION (multinuclear)

Definition: A PROPORTION relation expresses a proportionality or equivalence of tendency or degree between two nuclei. It is always multinuclear.

Example:

(272) "We can't have this kind of thing happen very often. [When the little guy gets frightened,] [the big guys hurt badly.] Merrill Lynch can't survive without the little guy."$_{wsj\_2386}$

41. PURPOSE (mononuclear)

Definition: In contrast to a RESULT relation, the situation presented in the satellite of a purpose relation is only putative, i.e., *it is yet to be achieved*. Most often it can be paraphrased as "nucleus in order to satellite."

Examples:

(273) [Bond Corp., a brewing, property, media and resources company, is selling many of its assets] [*to reduce its debts.*]$_{wsj\_0630}$

A purpose clause with a to-infinitive should not be confused with a postnominal modifier. For example, the text in italics below is not the satellite of a PURPOSE relation, but of an ELABORATION-OBJECT-ATTRIBUTE-E, since it modifies only the noun phrase *approval*, not the entire main clause. The unit in smaller fonts below is an embedded relative, not a purpose clause.

(274) [SA Brewing, an Australian brewer, last Thursday was given approval] *[to acquire an option for up to 20% of Bell Resources Ltd., a unit of Bond Corp.]*$_{wsj\_0630}$

42. QUESTION-ANSWER (multinuclear), QUESTION-ANSWER-N (mononuclear), QUESTION-ANSWER-S (mononuclear)

Definition: In a question-answer relation, one textual span poses a question (not necessarily realized as an interrogative sentence), and the other text span answers the question. The relation may be mononuclear or multinuclear, depending on the context. When the question is perceived as more important than the answer, the question is assigned the role of nucleus and the answer is the satellite. The relation QUESTION-ANSWER-S should be selected in this case. When the answer is the nucleus, use the label QUESTION-ANSWER-N; when the relation is multinuclear, use the relation QUESTION-ANSWER.

Mononuclear (QUESTION-ANSWER-N) Example:

(275) [*Asked about the speculation that Mr. Louis-Dreyfus has been hired to pave the way for a buy-out by the brothers,*] [the executive replied, "That isn't the

reason Dreyfus has been brought in.] [He was brought in to turn around the company.”]$_{wsj\_2331}$

Multinuclear Example:

(276) [But are these four players, three of them in their 80s, ready to assume a different role after 88 years, collectively, of service on the high court?] [Every indication is that the four are prepared to accept this new role, and the frustrations that go with it, but in different ways. Justices Brennan and Stevens appear philosophical about it; Justices Marshall and Blackmun appear fighting mad.]$_{wsj\_2347}$

43. REASON (both)

Definition: In a REASON relation, the nucleus must be an action carried out by an animate agent. Only animate agents can have reasons for performing actions. You can paraphrase it as “Satellite is the reason for Nucleus.” In cases where both spans appear equally important, select the multinuclear REASON.

Mononuclear Example:

(277) Earlier this year, DPC Acquisition made a $15-a-share offer for Dataproducts, [which the Dataproducts board said it rejected] *[because the $283.7 million offer was not fully financed.]*$_{wsj_0661}$

Multinuclear Example:

(278) Compare the past eight five-year plans with actual appropriations. [The Pentagon strategists produce budgets that simply cannot be executed] [because they assume a defense strategy that depends only on goals and threats. Strategy, however, is about possibilities, not hopes and dreams.]$_{wsj\_0692}$

44. RESTATEMENT (mononuclear)

Definition: A restatement relation is always mononuclear. The satellite and nucleus are of (roughly) comparable size. The satellite reiterates the information presented in the nucleus, typically with slightly different wording. It does not add to or interpret the information.

Example:

(279) “Once you add dramatizations, it's no longer news, it's drama, and that has no place on a network news broadcast... [They should never be on.] *[Never.”]*$_{wsj\_0633}$

45. RESULT (mononuclear)

Definition: The situation presented in the satellite is the cause of the situation presented in the nucleus. The result, which is the nucleus, is the most important part. Without presenting the satellite, the reader may not know what caused the result in the nucleus. In contrast to a PURPOSE relation, the situation presented in the nucleus of a result relation is factual, i.e., it is achieved. The intention of the writer is to emphasize the result. When the cause is the nucleus, select the mononuclear relation CAUSE. When it is not clear whether the cause or result is more important, select the multinuclear relation CAUSE-RESULT.

Example:

(280) [The explosions began] [*when a seal blew out.*]$_{wsj\_1320}$

46. RHETORICAL-QUESTION (mononuclear)

Definition: In a RHETORICAL-QUESTION relation, the satellite poses a question vis-a-vis a segment of the text; the intention of the author is usually not to answer it, but rather, to raise an issue for the reader to consider, or to raise an issue for which the answer should be obvious.

Examples:

(281) [So I said, ‘Hello.’ And she said, ‘Hello.’] [*Can you imagine?*] Liza said hello to me.$_{wsj\_1376}$

(282) [For the long-term investor who picks stocks carefully, the price volatility can provide welcome buying opportunities as short-term players scramble frantically to sell stocks in a matter of minutes.] [*Who can make the better decision, the guy who has 10 seconds to decide what to do or the guy with all the time in the world?*]$_{wsj\_0681}$

47. SAME-UNIT

Definition: A pseudo-relation used as a device for linking two discontinuous text fragments that are really a single EDU, but which are broken up by an embedded unit. Examples of embedded units that can break up other EDUs include: relative clauses, other nominal postmodifiers, parentheticals, participial clauses, etc. By convention, this relation is always multinuclear.

Example:

(283) [**The yen’s softness,**] [she says,] [**apparently stems from Japanese investors’ interest**] [in buying dollars against the yen to purchase U.S. bond issues] [**and persistent worries about this year’s upheaval in the Japanese government.**]$_{wsj\_0693}$

48. SEQUENCE

Definition: A SEQUENCE is a multinuclear list of events presented in chronological order. (See also INVERTED-SEQUENCE.)

Example: In the example below, there are three events that form a multinuclear SEQUENCE:

(284) [Ralph Brown was 31,000 feet over Minnesota when both jets on his Falcon 20 flamed out.] [At 18,000 feet, he says, he and his co-pilot "were looking for an interstate or a cornfield" to land.] [At 13,000 feet, the engines restarted.]$_{wsj\_1394}$

49. STATEMENT-RESPONSE (multinuclear), STATEMENT-RESPONSE-N (mononuclear), STATEMENT-RESPONSE-S (mononuclear)

Definition: In a STATEMENT-RESPONSE relation, one textual span presents a statement and the other span makes some sort of response to it. The statement may be one actually spoken by someone or the author's statement of a situation. Similarly, the response may be one actually spoken or a situational response to what is occurring in the statement portion. When the statement is perceived as more important than the response, the statement is assigned the role of nucleus and the response is the satellite. The relation STATEMENT-RESPONSE-S should be selected in this case. When the response is the nucleus, use the label STATEMENT-RESPONSE-N; when the relation is multinuclear, use the relation STATEMENT-RESPONSE.

Mononuclear (STATEMENT-RESPONSE-N) Example:

(285) [*Reports of Dr. Toseland's findings in the British press have triggered widespread concern among diabetics here.*] [Both the British Diabetic Association and the Committee on Safety in Medicines -- Britain's equivalent to the U.S. FDA -- recently issued statements noting the lack of hard scientific evidence to support Dr. Toseland's findings.]$_{wsj\_0690}$

Mononuclear (STATEMENT-RESPONSE-S) Example:

(286) [In a suit filed in federal court Thursday, the S&L alleged that a disproportionate number of the bonds it purchased in 1984 declined in value....] [*Officials at Drexel said they hadn't seen the suit and thus couldn't comment.*]$_{wsj\_0696}$

Multinuclear Example:

(287) [[Last week, Boeing Chairman Frank Shrontz sent striking workers a letter saying that "to my knowledge,

Boeing's offer represents the best overall three-year contract of any major U.S. industrial firm in recent history."] [But Mr. Baker called the letter -- and the company's offer of a 10% wage increase over the life of the pact, plus bonuses -- "very weak."]$_{wsj\_2308}$

50. SUMMARY-N (mononuclear), SUMMARY-S (mononuclear)

Definition: In a SUMMARY-S relation, the satellite summarizes the information presented in the nucleus. The emphasis is on the situation presented in the nucleus. The size of the summary (the satellite) is shorter than the size of the nucleus. In an SUMMARY-N relation, the nucleus summarizes the information presented in the satellite. The emphasis is on the summary. The size of the summary (the nucleus) is shorter than the size of the satellite.

Example of SUMMARY-N:

(288) [The Singapore and Kuala Lumpur stock exchanges are bracing for a turbulent separation, following Malaysian Finance Minister Daim-Zainuddin's long-awaited announcement that the exchanges will sever ties.] [*On Friday, Datuk Daim added spice to an otherwise unremarkable address on Malaysia's proposed budget for 1990 by ordering the Kuala Lumpur Stock Exchange to take appropriate action immediately to cut its links with the Stock Exchange of Singapore. The delisting of Malaysian-based companies from the Singapore exchange may not be a smooth process, analysts say. Though the split has long been expected, the exchanges aren't fully prepared to go their separate ways.*]$_{wsj\_0613}$

Example of SUMMARY-S:

(289) [The airline industry's fortunes, in dazzling shape for most of the year, have taken a sudden turn for the worse in the past few weeks. Citing rising fuel costs, promotional fare cuts and a general slowdown in travel, several major carriers have posted or are expected to post relatively poor third-quarter results....And they say the outlook for 1990 is nearly as bad.] [*Airlines in 1989 came in like a bang and are going out like a whimper, said Kevin Murphy, an airline analyst at Morgan Stanley & Co.*]$_{wsj\_1192}$

51. TEMPORAL-BEFORE (mononuclear)

Definition: In a TEMPORAL-BEFORE relation, the situation presented in the nucleus (often realized as a superordinate clause) occurs before or leading up to the situation in the satellite

(often realized as a subordinate clause). When the relation is multinuclear but the spans occur in reverse temporal order -- i.e., the situation presented in the second span occurs before the situation presented in the first span -- select the multinuclear relation INVERTED-SEQUENCE.

Example:

(290) [We want to make sure they know what they want] [*before they come back.*]$_{wsj\_2308}$

52. TEMPORAL-SAME-TIME (both)

Definition: In a TEMPORAL-SAME-TIME relation, the situations presented in the nucleus and satellite occur at approximately the same time, or at least there is an overlap between the two situations. This relation can be mononuclear or multinuclear.

Mononuclear Examples:

(291) [It's important for a new make to be as distinctive as possible] [*while still retaining links to the parent company's quality image.*]$_{wsj\_1377}$

(292) [One of those areas is the development of a hand-held electronic device that would permit floor traders to enter trades] [*as they make them.*]$_{wsj\_0664}$

Multinuclear Examples:

(293) [Prices of long-term treasury bonds moved inversely to the stock market] [as investors sought safety amid growing evidence the economy was weakening.]$_{wsj\_1151}$

(294) [By setting up the joint venture, Kaufman & Broad can take the more aggressive approach of buying raw land,] [while avoiding the negative impacts to its own balance sheet.]$_{wsj\_2313}$

53. TEMPORAL-AFTER (mononuclear)

Definition: In a TEMPORAL-AFTER relation, the situation presented in the nucleus (often realized as a superordinate clause) occurs after the situation presented in the satellite (often realized as a subordinate clause). When the relation is multinuclear, and the spans occur in temporal order -- i.e., the situation presented in the second segment occurs after the situation presented in the first segment -- select the multinuclear relation SEQUENCE.

Examples:

(295) [Small investors have tiptoed back into the market] [*following Black Monday.*]$_{wsj\_2386}$

(296) [RJR Nabisco Inc. is disbanding its division responsible for buying network advertising time,] [*just a month after moving 11 of the group's 14 employees to New York from Atlanta.*]$_{wsj\_2315}$

54. TEXTUAL-ORGANIZATION (multinuclear)

Definition: TEXTUAL-ORGANIZATION is a multinuclear relation used to link elements of the structure of the text, for example, to link a title with the body of the text, a section title with the text of a section, etc. The role of the relation is primarily that of enforcing a tree structure on the representation.

Examples:

(297) [Friday, October 13, 1989]^{Date} [The key U.S. and foreign annual interest rates below are a guide to general levels but don't always represent actual transactions....]$^{Text}_{wsj\_2380}$

(298) [Under a proposal by Democrats to expand individual retirement accounts, a $2,000 contribution by a taxpayer in the 33% bracket would save $330 on his taxes. The savings was given incorrectly in Friday's edition.]^{Text} [See: Politics and Policy: Debate on IRAs Centers on Whether Tax Break Should Be Immediate or Put Off Till Retirement -- WSJ Oct. 27, 1989)]$^{Footnote}_{wsj_0605}$

55. TOPIC-COMMENT (multinuclear)

Definition: A general statement or topic of discussion is introduced, after which a specific remark is made on the statement or topic. This relation is always multinuclear, as both spans are necessary to understand the context. When the spans occur in the reverse order, with the comment preceding the topic, the relation COMMENT-TOPIC is selected.

Example:

(299) [As far as the pound goes,] [some traders say a slide toward support at $1.5500 may be a favorable development for the dollar this week.]$_{wsj_0693}$

56. TOPIC-DRIFT (both)

Definition: The relation TOPIC-DRIFT is used to link large textual spans when the topic drifts smoothly from the information presented in the first span to the information presented in the second. The same elements are in focus in both textual units. While this relation may be either mononuclear or multinuclear, it is usually multinuclear. Only select mononuclear if the relative size or importance of one of the spans is less significant than that of the other.

Multinuclear Example:

(300) [Food and Drug Administration spokesman Jeff Nesbit said the agency has turned over evidence in a criminal investigation concerning Vitarine Pharmaceuticals Inc....] [Mr. Nesbit also said the FDA has asked Bolar Pharmaceutical Co. to recall at the retail level its urinary tract antibiotic....]$_{wsj\_2382}$

57. TOPIC-SHIFT (both)

Definition: The relation TOPIC-SHIFT is used to link large textual spans when there is a sharp change in focus going from one segment to the other. The same elements are NOT in focus in the two spans. While this relation may be either mononuclear or multinuclear, it is usually multinuclear. Only select mononuclear if the relative size or importance of one of the spans is less significant than that of the other.

Multinuclear Example:

(301) [South Africa freed the ANC's Sisulu and seven other political prisoners.... Mandela, considered the most prominent leader of the ANC, remains in prison. But his release within the next few months is widely expected.] [The Soviet Union reported that thousands of goods needed to ease widespread shortages across the nation were piled up at ports and rail depots, and food shipments were rotting because of a lack of people and equipment to move the cargo.]$_{wsj\_2356}$

II.2. Schemata

Schemata are labels that refer to structural elements of the organization of a text, such as a title, author, signature block, or the body of the text itself. They are associated with individual nodes in the discourse structure of a text, and represent an annotation level that is independent of the rhetorical relations defined above. Schemata do not reflect relations between text spans as rhetorical relations do, but rather characterize a functional role of an individual text span. In terms of the physical layout of a text, a schema pertains to a block or segment of the text.

Schemata are linked to other schemata by the multinuclear relation TEXTUALORGANIZATION. In the examples below, a schema label is shown as a superscript at the end of a bracketed unit to which it applies. The corresponding bracketed unit and schema to which it is linked via the TEXTUALORGANIZATION relation is also shown.

1. AUTHOR

Definition: Most of the time, the WSJ articles in our corpus did not include title and author information. However, in a small number of cases, a line about the author was included at the end of the text. In such cases, the schema Author was used to link this with the rest of the text.

Example:

(302) [When the Trinity Repertory Theater named Anne Bogart its artistic director last spring, the nation's theatrical cognoscenti arched a collective eyebrow. ... But it is the Trinity Rep newcomer, Jonathan Fried (Zamislov, the paralegal) who is the actor to watch, whether he is hamming it up while conducting the chamber musicians or seducing his neighbor's wife (Becca Lish) by licking her bosom.]Text [Ms. de Vries writes frequently about theater.]Author wsj_1163

2. ABSTRACT

Definition: When a separate section of a document contains an abstract or summary of the document, this schema should be used. If, within the text of the article, a sentence or paragraph adequately summarizes the document, then the relation SUMMARY-S or SUMMARY-N should be selected. The schema ABSTRACT is reserved for a summary which is separate from the body of the text itself.

Examples: There were no examples in this corpus.

3. COLUMN-TITLE

Definition: This schema refers to the name of a regularly featured column in the paper, and should be used when the paper begins by presenting a TITLE, COLUMN-TITLE and AUTHOR.

Examples: There were no examples in this corpus, since the articles typically began with the body of the text. However, in wsj_1107, reference to a column title (underlined) is made in the following sentence, which also refers to the author and the title for that particular day's column: *This is in response to George Melloan's Business World column "The Housing Market Is a Bigger Mess Than You Think" (op-ed page, Sept. 26).*

4. DATE

Definition: When a date is provided at the beginning of the document, and it refers to the rest of the article, the DATE schema should be used. It may also be used for subsections of the document. It should be linked to a TEXT or SECTIONTEXT schema.

Example:

(303) [Friday, October 27, 1989]Date [The key U.S. and foreign annual interest rates below are a guide to general levels but don't always represent actual transactions. ...]Text wsj_0602

5. FOOTNOTE

Definition: Supplementary information given outside of the body of the text, usually at the end of a document in this corpus.

Example:

(304) [THE FINANCIAL ACCOUNTING STANDARDS BOARD'S coming rule on disclosure involving financial instruments will be effective for financial statements with fiscal years ending after June 15, 1990. The date was misstated in Friday's edition.]^{Text} [(See: “FASB Plans Rule on Financial Risk of Instruments” -- WSJ Oct. 27, 1989)]$^{Footnote}_{wsj\_0603}$

6. HEADING

Definition: The WSJ articles in our corpus did not include headings, so this schema was not used in tagging the corpus.

Examples: none

7. POINT-OF-ORIGIN

Definition: POINT-OF-ORIGIN refers to the geographic location in which the events in the story take place. Use this schema if this information is designated separately at the beginning or end of the article or section of the article. It should be linked to a TEXT or SECTIONTEXT schema.

Example:

(305) [CHICAGO:]$^{Point\text{-}of\text{-}Origin}$ [Sears, Roebuck & Co. is struggling as it enters the critical Christmas season....]$^{Text}_{wsj\_1105}$

8. SECTIONTEXT

Definition: Often, an article will be broken up structurally into distinct sections representing different topics. In this case, the schema SectionText is used. Most often, each section will be labeled, so there will be a corresponding SECTIONTITLE and SECTIONTEXT.

Examples:

(306) [Exxon]^{SectionTitle}
[Although Exxon spent heavily during the latest quarter to clean up the Alaskan shoreline blackened by its huge oil spill, those expenses as well as the cost of a continuing spill-related program are covered by $880 million in charges taken during the first half...]^{SectionText}

```
[Ashland Oil]SectionTitle
[A rash of one-time charges left Ashland Oil with a loss
of $39 million for its fiscal fourth quarter. A year
earlier, the refiner earned $66 million, or $1.19 a
share. Quarterly revenue rose 4.5%, to $2.3 billion
from $2.2 billion. For the year, net income tumbled 61%
to $86 million, or $1.55 a share...]SectionText
[Amerada Hess]SectionTitle
[Third-quarter earnings at Amerada Hess more than tri-
pled to $51.81 million, or 64 cents a share, from $15.7
million, or 20 cents a share, a year earlier. Revenue
climbed 28%, to $1.18 billion from $925 mil-
lion.]SectionText wsj_1311
```

9. SECTIONTITLE

Definition: Often, an article will be broken up structurally into distinct sections representing different topics. Most often, each section will be labeled, so there will be a corresponding SECTIONTITLE and SECTIONTEXT. Usually, the section title occurs on a separate line.

Examples: See above under SECTIONTEXT.

10. SIGNATURE-BLOCK

Definition: A SIGNATURE-BLOCK is a section of the document that may contain a person's name, title and company. If these occur on multiple lines, each line should be treated as a separate EDU, and all should be linked via a LIST relation before the SIGNATURE-BLOCK label is selected. This occurs in letters to the editor in the WSJ corpus. This schema is linked to the schema TEXT (if the entire article is a letter) or SECTIONTEXT (if only a portion of the article is a letter).

Example:

```
 (307)  [The Federal Housing Administration, Veterans
Administration and the Department of Housing and Urban
Development further aggravate the problem of affordable
housing stock by “buying in” to their foreclosed prop-
erties (of which there are, alas, many) at an inflated
“balance due” -- say $80,000 on a house worth $55,000 -
- instead of allowing a free market to price the house
for what it's really worth. Worse, the properties then
sit around deteriorating for maybe a year or so, but are
resold eventually (because of the attractiveness of the
low down payment, etc.) to a marginal buyer who can't
afford both the mortgage and needed repairs; and having
little vested interest that buyer will walk away and
```

the vicious cycle repeats itself all over again.]Section-Text [Paul Padget Cincinnati]Signature-Block wsj_1107

11. SOURCE

Definition: The schema SOURCE is used when an article, or section of an article, begins or ends by noting the source of the information presented in the article. It should be linked to a TEXT or SECTIONTEXT schema.

Example:

(308) [From the Sept. 30-Oct. 4 issue of The Economist:]Source [What defeated General Aoun was not only the weight of the Syrian army. ...]Text wsj_1141

12. TEXT

Definition: The schema Text is used to link the main body of the text to other structural elements of the document, such as a FOOTNOTE, SOURCE, HEADING, etc. This schema is used only if the entire body of the text should be linked to another schema. If only a subsection of the text is relevant, the schema SECTIONTEXT is used instead. Occasionally, a WSJ document contains more than one article. In that case, the schema TEXT may be used, for example to link a TITLE and TEXT for each of the three articles included within a single document (see example below under 13. Title).

Examples: (These are repeated from other schemata above.)

(309) [CHICAGO:]Point-of-Origin [Sears, Roebuck & Co. is struggling as it enters the critical Christmas season. ...]Text wsj_1105

(310) [Friday, October 27, 1989]Date [The key U.S. and foreign annual interest rates below are a guide to general levels but don't always represent actual transactions. ...]Text wsj_0602

13. TITLE

Definition: The WSJ documents from the LDC usually did not include titles. However, in a small number of cases, a WSJ document consisted of more than one article, each of which included a title. In that case the schemata TITLE and TEXT were used.

Examples:

(311) [Good grief! Charlie Brown is selling out.]Title [Those Metropolitan Life ads were bad enough. But now, Charlie Brown is about to start pitching everything

from Chex Party Mix to light bulbs. ...]^{Text} [Berry Rejoins WPP Group]^{Title} [Norman Berry, the creative executive who was apparently squeezed out of Ogilvy & Mather in June, is returning to Ogilvy's parent company, WPP Group PLC. ...]^{Text} [RJR Taps FCB/Leber]^{Title} [RJR Nabisco Inc. awarded its national broadcast media-buying assignment to FCB/Leber Katz Partners, the New York outpost of Chicago-based Foote, Cone & Belding. ...]^{Text} wsj_1193

Appendix III: Cue Phrases

This appendix provides an alphabetical listing of sample cue words and phrases that may indicate a rhetorical relation. It is intended as a guide, and not as an exhaustive listing. Examples illustrate typical rhetorical relations these cues may appear in. As elsewhere in this manual, only the EDUs relevant to the discussion are marked.

1. after

TEMPORAL-AFTER relation:

(312) [The dollar finished lower yesterday] [***after*** *tracking another rollercoaster session on Wall Street.*]$_{wsj\_1102}$

CIRCUMSTANCE relation:

(313) [***After*** *its previous mayor committed suicide last year,*] [an investigation disclosed that town officials regularly voted on their own projects, gave special favors to developer friends and dipped into the town's coffers for trips and retreats.]$_{wsj\_2367}$

2. although

CONCESSION relation:

(314) [***Although*** *they represent only 2% of the population,*] [they control nearly one-third of the discretionary income.]$_{wsj\_2366}$

ANTITHESIS relation:

(315) [***Although*** *external events have contributed to the morass,*] [the principal causes of the current crisis are internal and generic to all programs.]$_{wsj\_1131}$

3. as

CIRCUMSTANCE relation:

(316) [The agreement was announced by Costa Rican President Oscar Arias Friday,] [***as*** *President Bush and other leaders from the Western Hemisphere gathered in the Central American nation for a celebration of democracy.*]$_{wsj\_0924}$

TEMPORAL-SAME-TIME relation

(317) [One of those areas is the development of a hand-held electronic device that would permit floor traders to enter trades] [***as*** *they make them.*]$_{wsj\_0664}$

4. as long as

CONDITION relation:

(318) [Encourage long-term occupancy by forgiving one month's payment (off the tail end of the mortgage) for every six months paid; or perhaps have the down payment deferred to the end of the mortgage (balloon), but for-given on a monthly pro-rata basis] [***as long as*** *the owner remains the occupant.*]$_{wsj\_1107}$

5. because

EXPLANATION-ARGUMENTATIVE relation:

(319) [Those bills can't be vetoed in their entirety] [***because*** *they are often needed to keep the government operating.*]$_{wsj\_0609}$

REASON relation:

(320) [Our research shows we sell more of our heavier issues] [***because*** *readers believe they are getting more for what they pay for.*]$_{wsj\_1123}$

6. by

MEANS relation:

(321) [Maybe she could step across the Plaza to the Met -- where she has still to make a debut -- and help out her Czech compatriot] [***by*** *singing the slow parts of Traviata.*]$_{wsj\_1154}$

7. despite

CONCESSION relation:

(322) [***Despite*** *their considerable incomes and assets,*] [40% of the respondents in the study don't feel financially secure, and one-fourth don't feel that they have made it.]$_{wsj\_2366}$

ANTITHESIS relation:

(323) [***Despite*** *the inevitable comparison with Compaq, however,*] [Texas Instruments' new notebook won't be a direct competitor.]$_{wsj\_0638}$

8. following

CIRCUMSTANCE relation:

(324) [Prime Computer plans to dismiss 20% of its work force to cut costs] [***following*** *its recent leveraged buy-out.*]$_{wsj\_1364}$

TEMPORAL-AFTER relation:

(325) [Small investors have tiptoed back into the market] [***following*** *Black Monday.*]$_{wsj\_2386}$

9. however

ANTITHESIS relation:

(326) [General Motors, for example, uses metric terms for its automobile bodies and power trains.] [*In auto advertising,* ***however****, items such as wheelbases are still described in inches*.]$_{wsj\_0676}$

CONTRAST relation:

(327) [The small changes in average reflect generally unchanged yields at many major banks.] [Some, **however**, lowered yields significantly.]$_{wsj\_1166}$

10. if

CONDITION relation:

(328) [***If*** *HDTV takes off in the U.S.*] [there will be demand for some 4,000 to 5,000 HDTV converters, known in the industry as telecines.]$_{wsj\_1386}$

11. meanwhile

TEMPORAL-SAME-TIME relation:

(329) [Thus, optimistic entrepreneurs await a promised land of less red tape -- just as soon as Uncle Sam gets around to arranging it.] [**Meanwhile**, they tackle the mounds of paper -- and fantasize about a dream world

where bulk-mail postal regulations and government inspectors are banished.]$_{wsj\_1162}$

TOPIC DRIFT relation:

(330) [Despite several spurts of dollar trading, it was noted that mark-yen cross trade grabbed much of the market's attention.... Despite the yen's weakness with respect to the mark, Tokyo traders say they don't expect the Bank of Japan to take any action to support the Japanese currency on that front.] [**Meanwhile**, sterling slumped on news that the United Kingdom posted a wider-than-expected deficit in December.]$_{wsj\_1102}$

12. since

TEMPORAL-AFTER relation:

(331) [The project would represent the single largest investment in the Philippines] [***since*** *President Corazon Aquino took office in February 1986.*]$_{wsj\_0606}$

CIRCUMSTANCE relation:

(332) [***Since*** *founding the company,*] [the charismatic Vietnam vet, who is still only 46 years old, has fostered an ethos of combat.]$_{wsj\_1394}$

13. so

CONSEQUENCE-S relation:

(333) [It's written 'marcato' in the score, and I played it that way, kind of gigue-like. And he yelled out 'dolce! dolce!' {sweet! sweet!}] [***So*** *we did it over, he adds.*]$_{wsj\_1388}$

PURPOSE relation:

(334) [Indeed, during a recent post-production audience discussion, the director explained that her fondest artistic wish was to find a way to play Somewhere Over the Rainbow] [***so*** *that the song's original beauty comes through, surmounting the cliche.*]$_{wsj\_1163}$

14. until

TEMPORAL-BEFORE relation:

(335) [***Until*** *Mr. Luzon took the helm last November*] [Banco Exterior was run by politicians who lacked either the skills or the will to introduce innovative changes.]$_{wsj\_0616}$

CONDITION relation:

(336) [In the same month, the Office of Thrift Supervision ordered the institution to stop paying common stock dividends] [***until*** *its operations were on track.*]$_{wsj\_2360}$

15. when

CIRCUMSTANCE relation:

(337) [Where the hell are they gonna live] [***when*** *people like you turn the world into a big toxic dump?*]$_{wsj\_1907}$

TEMPORAL-SAME-TIME relation:

(338) [Ralph Brown was 31,000 feet over Minnesota] [***when*** *both jets on his Falcon 20 flamed out.*]$_{wsj\_1394}$

16. while

TEMPORAL-SAME-TIME relation:

(339) [Like Peter Sellars, Ms. Bogart manipulates her actors as if they were rag dolls, sprawling them on staircases, dangling them off table, even hanging them from precipices] [***while*** *having them perform some gymnastic feats of derring-do.*]$_{wsj\_1163}$

CIRCUMSTANCE relation:

(340) [***While*** *the student was in school,*] [interest costs would either be paid by the student or added to the loan balance.]$_{wsj\_1131}$

ANTITHESIS relation:

(341) [***While*** *two-thirds feel some guilt about being affluent*] [only 25% give $2,500 or more to charity each year]$_{wsj_2366}$

COMPARISON relation:

(342) [Kellogg's current share is believed to be slightly under 40%] [***while*** *General Mills' share is about 27%.*]$_{wsj\_0610}$

CONTRAST relation:

(343) [But the staff at some of those locations will be slashed] [**while** at other the work force will be increased.]$_{wsj\_0688}$

17. without

MANNER relation:

(344) [A judge must jump from murder to antitrust cases, from arson to securities fraud] [***without*** *missing a beat.*]$_{wsj\_0601}$

CIRCUMSTANCE relation:

(345) [***Without*** *many actual deals to show off,*] [Kidder is left to stress that it finally has a team in place, and that everyone works harder.]$_{wsj\_0604}$

Appendix IV: Using the Discourse Annotation Tool

The Discourse Annotation Tool enables you to build incrementally the discourse structures of texts. The Interface to the Annotation Tool has two distinct panels:

• The discourse structure (RST) panel (the one at the top) allows you to construct/modify/ delete the discourse structure of a text.

• The text panel (the one at the bottom) allows you to read the text of interest incrementally (one sentence at a time) and to mark the elementary units of discourse.

Figure XX presents a snapshot of the tool.

IV.1. How to Annotate Texts

In annotating the discourse structure of a text, you should follow these steps:

1. Start the tool using the command <home directory>/RSTTool/RSTTool.
2. Use the **File/Load Text File** menu in order to load a text into the annotation tool.

When a text is loaded, the text panel shows the first sentence of the text and the RST panel is initialized to an empty discourse structure. If you have annotated the text previously, you can either choose to start the annotation process from scratch or you can re-load the part of the text that has been already annotated. The latter option allows you to interrupt and resume the annotation process at any desired time.

3. Once a text is loaded, you should follow repeatedly these steps:

(a) Use the left button of your mouse in order to mark in the text panel the right boundary of an elementary unit.

The boundaries of each elementary unit are given by two consecutive markers, which are automatically assigned natural number labels from <1> to <*n*>, where *n* is the total number of elementary units in a text. The first marker <0> that corresponds to the beginning of the text is not shown.

The identification of a boundary automatically creates in the RST panel an elementary textual span. If you create a boundary at the end of a sentence, the next sentence will be automatically displayed in the text panel. Hence you can see only one sentence at a time.

(b) Link the new elementary unit to a simple or complex span that was created before by clicking the left button of your mouse on the span where you want the new unit to be attached. When you click on the span where the new unit is going to be attached, you are supposed to use a pop-up menu in order to specify the type of relation that holds between the two units. The relation is:

- ***satellite***, if the new unit is the satellite of a mononuclear relation.

• ***nucleus***, if the new unit is the nucleus of a mononuclear relation.

• ***multinuclear***, if both units have a nuclear status.

• ***embedded-satellite***, if the new unit is embedded into the unit to which you attach it.

• ***embedded-nucleus***, if the unit/span to which the new unit is going to be linked is embedded into the new unit.

• ***schema***, if the new unit is the same as the old unit and plays a predetermined role in the text, such as title, author, section, etc.

Once you assign the nuclearity status, you may then select the type of relation according to the relation definitions that are provided in this document.

4. When all elementary units of a text are identified and all units in the discourse structure panel are linked in a hierarchical structure that conforms to the requirements of RST, you can exit or you can load another text file.

IV.2. Notes on the Annotation Process

• The annotation tool implements a very simple algorithm for identifying sentence ends. As a consequence, it is possible that a sequence of words displayed in the text panel is not a full sentence. When this situation occurs, you can use the **Next sentence** button in order to see more text.

• In some cases, it is impossible for you to determine where and how a new textual unit is to be incorporated in the discourse structure that you have already built. This is normal. Don't panic. Identify the boundaries of the discourse units that follow. When you know how to link the units, do it by clicking on the **Link** button. When you are in “Link” mode (and not in “Auto” mode, which is the default mode) you have to click first on one of the discourse segments that you want to link, and without releasing the button drag the mouse to the other segment that will participate in the relation. Release then the button. As a result of this operation, you will have to specify through a menu-driven dialogue the type of relation and the rhetorical status of the segment on which you have clicked first (satellite or nucleus if the relation is mononuclear; multinuclear if the relation is multinuclear; embedded-satellite or embedded-nucleus if one of the units/spans is embedded in the other unit/span.

Warning: Make sure that you only link adjacent spans!!! If you don't do so, the resulting structure will no longer be a well-formed discourse tree! After you have linked two segments, the annotation tool switches automatically to the “Auto mode”.

• Whenever you make a mistake, you can use the **Undo** button to go back as many steps as you want. One click of the **Undo** button will cancel the last discourse operation in the RST panel. Another click will cancel the last discourse boundary inserted in the text panel.

• At any time, you can change the type of the relation that is associated with two spans by clicking on the **Change relation** button and then by clicking on the satellite of that relation if the relation is mononuclear, or on one of its nuclei if the relation is multinuclear.

Warning 1: You can change in this way the relation name only! You cannot change the status of a discourse segment. If you want to do that, you will have to

use the **Modify** or **Disconnect** buttons, or **Undo** as many steps of your annotation process as is necessary.

Warning 2: In this way, you can replace a relation name only with a relation from the same group. For example, if a relation was originally assumed to be multinuclear, the **Change relation** button will allow you to choose only another type of multinuclear relation. If you want to change the type of the relation (from satellite to embedded-satellite, for example), you will have to use the **Modify** or **Disconnect** buttons, or "Undo" as many steps of your annotation process as is necessary.

• If you want to change the nuclearity or type of a relation or if you want to modify the structure you've built, you should use the **Modify** button. After you click on **Modify**, you can click on any satellite of a mononuclear relation or on any nucleus of a multinuclear rhetorical relation. As a result, the span on which you've clicked will be disconnected from the tree. In the next immediate step, you are supposed to click on the node where you want the disconnected span to be reattached. When you do so, you are taken through the same menu as in the case of the **Link** button. You can use this to change completely the shape of the tree that you build or to simply change the polarity of a relation. For example, if unit 1 is the satellite of unit 2 and you want to change the polarity of the relation, you will have to:

1) click on the modify button;

2) click on node 1, which is the satellite; As a result, this node will be disconnected from its nucleus.

3) click on node 2 and specify that the relation is of type "nucleus". As a result, node 1 will be assigned status "nucleus", and node 2 status "satellite". The relation will be the one you have chosen.

Warning: In some cases, the "Disconnect" and "Modify" actions create/destroy existing nodes in a way that can introduce errors in the log files that reflect all the actions you have taken (some of the relations may point to nodes that have been destroyed). If you plan to use the log files extensively, try to avoid using these buttons when you disconnect internal spans of trees. If you plan to use only the final structures that you build, you don't have to worry about this issue.

• The "Disconnect" button allows you to disconnect a subtree. Use it only when you want to re-shape the tree that you build through a succession of complex "Disconnect" and "Link" operations. When you know where the disconnected tree is to be attached, use the **Modify** button.

• During the annotation process, as texts get larger, it is impossible to visualize the whole tree on the screen. By clicking the middle button of your mouse on a node of the discourse structure that you have built, that node is collapsed (the children nodes become invisible, and so the tree shrinks). By clicking the right button of your mouse on a collapsed node, that node is expanded to its original size.

• The annotation tool attempts to keep in focus the last unit that was added to the discourse structure. If the text displayed in the RST panel is truncated, you can use the **Enlarge** button to

increase the size of the RST panel. The **Reduce** button can be used for the reverse operation. However, in most cases, you will notice that by simply using the slide rules you can focus on the subtrees of interest. In fact, when you are in “Auto” mode, it is enough if you focus on the node where you want the new node to be linked (for convenience, the last selected unit is always displayed between the two panels). In “Link” mode, you will have to make sure that both nodes that you intend to link are visible at the same time. If they are not, collapse them first.

• Some textual units play multiple roles in the text. For example, unit 7 is both in a List relation with units [5,6] and an Elaboration-Set-Member relation with unit 4. If we attach unit 7 directly to unit 4 through an Elaboration-Set-Member relation, we do not capture the LIST relation with units [5,6]. However, if we attach unit 7 through a LIST relation to units [5,6], we also capture implicitly the ELABORATION-SET-MEMBER relation.

Whenever you build a discourse structure, choose a representation that explicitly and implicitly represent as many relations as possible!

```
[Smart cards are becoming more attractive2][as the
price of microcomputing power and storage continues to
drop.3] [They have two main advantages over magnetic-
stripe cards.4][First, they can carry 10 or even 100
times as much information5][- and hold it much more
robustly.6][Second, they can execute complex tasks in
conjunction with a terminal.7]
```

• During the annotation process, always follow these rules:

1) Whenever you are in “Link” mode, connect the last two segments in the discourse representation as much as such an approach enables you to build a tree that is correct.

2) Make sure the final structure of a text has only one root node.

3) If you cannot choose between a multinuclear and a mononuclear relation, prefer the mononuclear one.

• In our initial experiments with the tool, we tried to constrain annotators to build the trees incrementally, to mirror current models of discourse parsing. We found that this procedure did not work very well. Quite often, it was unclear how to link a unit to the preceding discourse.

• Some annotators found that printing and reading the text before annotating it helped. (If you just want to read the text, you can always choose the **Show Text** button.) Others found that determining the large textual segments beforehand is useful too.

• If obtaining good trees is your top priority and you don't care about being consistent with a parsing model or other, build the trees whatever way is more comfortable for you.

Appendix V: References

Carlson, Lynn; Marcu, Daniel; and Okurowski, Mary Ellen. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialog,* Aalborg, Denmark.

Ferrari, Giacomo. 1998. Preliminary steps toward the creation of a discourse and text resource. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998),* Granada, Spain, 999-1001.

Garside, Roger; Fligelstone, Steve; and Botley, Simon. 1997. Discourse Annotation: Anaphoric Relations in Corpora. In *Corpus Annotation: Linguistic Information from Computer Text Corpora,* edited by R. Garside, G. Leech, and T. McEnery. London: Longman, 66-84.

Grosz, Barbara and Sidner, Candice. 1986. Attentions, intentions, and the structure of discourse. *Computational Linguistics,* 12(3): 175-204.

Leech, Geoffrey; McEnery, Tony; and Wynne, Martin. 1997. Further Levels of Annotation. In *Corpus Annotation: Linguistic Information from Computer Text Corpora,* edited by R. Garside, G. Leech, and T. McEnery. London: Longman, 85-101.

Mann, William and Thompson, Sandra. 1988. Rhetorical structure theory. Toward a functional theory of text organization. *Text,* 8(3): 243-281.

Marcu Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA: The MIT Press.

Marcu, Daniel; Amorrortu, Estibaliz; and Romera, Magdalena. 1999. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging,* College Park, MD, 48-57.

Marcus, Mitchell; Santorini, Beatrice; and Marcinkiewicz, Mary Ann. 1993. Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics* 19(2), 313-330.

Passonneau, Rebecca and Litman, Diane. 1997. Discourse segmentation by human and automatic means. Computational Linguistics 23(1).: 103-140.

Van Dijk, Teun A. and Kintsch, Walter. 1983. *Strategies of Discourse Comprehension*. New York: Academic Press.