RST Signalling Corpus Annotation Manual

Debopam Das and Maite Taboada
Department of Linguistics
Simon Fraser University
ddas@sfu.ca, mtaboada@sfu.ca

September 17, 2014

Contents

1. Introduction

This manual documents the guidelines used in annotating the signals of coherence (rhetorical) relations in the RST Discourse Treebank (Carlson et al., 2002). The RST Discourse Treebank (RST-DT) includes a collection of newspaper texts already annotated for coherence relations. In the RST Signalling Corpus, we have added a new layer of signalling information to the existing RST-DT with the aim to find out how coherence relations are signalled in discourse. More information about the annotation project can be found in Debopam Das' PhD dissertation "Signalling of Coherence Relations in Discourse" completed at Simon Fraser University (SFU) in Summer 2014. The dissertation is available through the SFU library (http://www.lib.sfu.ca/help/publication-types/finding-sfu-theses), and the RST Signalling Corpus is available through the Linguistic Data Consortium or LDC (https://www.ldc.upenn.edu/).

In this manual, we provide a brief description of the RST-DT, the annotation process, and the description (classification, definition and examples) of the signals used for annotation.

2. RST Discourse Treebank

2.1. RST-DT: An Overview

The RST Discourse Treebank or RST-DT (Carlson et al., 2002) is a corpus annotated for coherence relations, and it contains a collection of 385 Wall Street Journal articles (representing over 176,000 words of text) selected from the Penn Treebank (Marcus et al., 1993). The corpus is distributed by the Linguistic Data Consortium, from which the corpus can be downloaded (for a fee).

The articles chosen for annotation in the RST-DT come from a variety of topics, such as financial reports, general interest stories, business-related news, cultural reviews, editorials, and letters to the editor (Carlson et al., 2001). The texts in these articles are annotated manually by a group of annotators. The annotation process is aided by the use of a tool, a modified version of the RSTTool (O'Donnell, 1997). The tool provides a graphical representation of the discourse structures of a text in the form of tree-diagrams, and it stores the annotated texts as LISP files. The reliability of the annotations is measured for four levels: elementary discourse units, hierarchical spans, hierarchical nuclearity and hierarchical relation assignments. The results of the inter-annotator agreement show considerably higher scores in all these four levels of annotations[1].

2.2. Theoretical Framework: RST

The relational annotation in the RST-DT is performed following Rhetorical Structure Theory or RST (Mann & Thompson, 1988). Carlson et al. (2001) find three reasons for using RST as the theoretical framework for their annotation work. The reasons are: (i) RST produces rich annotations that uniformly represent intentional, semantic and textual features of texts; (ii) discourse annotations by multiple judges within the RST framework yield relatively higher levels of agreement; and (iii) the use of RST trees prove to be beneficial in many NLP applications such as natural language generation, text summarization, machine translation and essay-scoring systems; and this suggests that RST can also be applied for other NLP-related resources, such as in the building of a discourse annotated treebank.

[1] See Carlson et al. (2001) for more detail.

Carlson et al. (2001) use a modified version of the RST framework in their annotation scheme (Carlson & Marcu, 2001). These modifications include some refinements in the treatment of elementary discourse units and in the taxonomy of RST relations used for the annotation.

2.2.1. Elementary Discourse Unit

Carlson and Marcu (2001) generally consider clauses to be the minimal units of discourse. However, they specify certain conditions for a clause to be used as an elementary discourse unit (EDU). The conditions are enumerated with examples below[2].

- Clauses that are subjects or objects of a main verb are not considered to be EDUs.

  (1) [Deciding what constitute "terrorism" can be a legalistic exercise.] wsj_1101

  (2) [Making computer smaller often means sacrificing memory.] wsj_2387

- Clauses that are complements of a main verb are not considered to be EDUs.

  (3) [With the golden share as protection, Jaguar officials have rebuffed Ford's overtures, and moved instead to forge an alliance with GM.] wsj_0632

  (4) [The company's current management found itself "locked into this," he said.] wsj_1103

- Complements of attribution verbs (speech acts and other cognitive acts) are considered to be EDUs.

  (5) [The legendary GM chairman declared] [that his company would make "a car for every purse and purpose".] wsj_1377

  (6) [Analysts estimated] [that sales at U.S. stores declined in the quarter, too.] wsj_1105

- Relative clauses, nominal postmodifiers, or clauses that break up other legitimate EDUs, are considered to be embedded discourse units.

  (7) [Some entrepreneur say] [the red tape] [they most love to hate] [is red tape] [they would also hate to lose.] wsj_1162

[2] The texts in the examples are taken from the RST-DT (Carlson et al., 2002). The text within square brackets denotes a span. A span (or a sequence of spans) is followed by a file number, denoting the article (in the RST-DT) from which the text is taken. For highlighting a particular clause, the relevant parts in a span are underlined.

(8) [The fact] [that this happened two years ago] [and there was a recovery] [gives people comfort] [that this won't be a problem.] wsj_2345

- Phrases that begin with a strong DM, such as *because, despite, without* and *as a result* are considered to be EDUs.

(9) [Today, no one gets in or out of the restricted area] [without De Beer's stingy approval] wsj_1121

(10) [Some big brokerage firms said] [they don't expect major problems] [as a result of margin calls.] wsj_2393

2.2.2. Taxonomy of Relations

Carlson et al. (2001), for the development of the RST-DT, use a large set 78 relations which are divided into 16 major relation groups (Carlson & Marcu, 2001). Carlson et al.'s taxonomy of RST relations is provided in Table 1.

#	Relation Group	Relation
1.	Attribution	Attribution, Attribution-negative
2.	Background	Background, Circumstance
3.	Cause	Cause, Result, Consequence
4.	Comparison	Comparison, Preference, Analogy, Proportion
5.	Condition	Condition, Hypothetical, Contingency, Otherwise
6.	Contrast	Contrast, Concession, Antithesis
7.	Elaboration	Elaboration-additional, Elaboration-general-specific, Elaboration-part-whole, Elaboration-process-step, Elaboration-object-attribute, Elaboration-set-member, Example, Definition
8.	Enablement	Purpose, Enablement
9.	Evaluation	Evaluation, Interpretation, Conclusion, Comment
10.	Explanation	Evidence, Explanation-argumentative, Reason
11.	Joint	List, Disjunction
12.	Manner-Means	Manner, Means
13.	Topic-Comment	Problem-solution, Question-answer, Statement-response, Topic-comment, Comment-topic, Rhetorical-question
14.	Summary	Summary, Restatement
15.	Temporal	Temporal-before, Temporal-after, Temporal-same-time, Sequence, Inverted-sequence
16.	Topic Change	Topic-shift, Topic-drift

Table 1: Taxonomy of RST relations in the RST-DT

Note that the total number of individual relations in Table 1 is less than 78. This is because only the basic relations are listed in this taxonomy, irrespective of being further categorized with respect to

nuclearity status (mononuclear/multinuclear) or the presence of relevant relational content in the nucleus or satellite. A single relation (such as *Cause* or *Result*) can be used as both mononuclear and multinuclear, based on the relative importance of the spans. Furthermore, a mononuclear/multinuclear relation such as *Evaluation* can be divided into *evaluation-n* or *evaluation-s*, implying the presence of the relevant relational content in the nucleus or satellite, respectively. The complete taxonomy, including all 78 relations, is provided in Appendix A.

Furthermore, three additional relations: *Textual-Organization*, *Span* and *Same-Unit*, have been used in the annotation of the RST-DT to in order to impose certain structure-specific requirements on the discourse trees.

3. Signals of Coherence Relations

The most important aspect of the signalling annotation task is to select and classify the types of signals to annotate. We built our taxonomy of signals based on the different classes of relational markers that have been mentioned in previous studies, or that we identified in our preliminary corpus work (Das, 2012; Das & Taboada, 2013; Taboada & Das, 2013).

The taxonomy of signals used in our annotation is organized hierarchically in three levels: *signal class*, *signal type* and *specific signal*. The top level, *signal class*, has three tags representing three major classes of signals: *single*, *combined* and *unsure*. For each class, a second level is defined; for example, the class *single* is divided into nine types (*DMs*, *reference*, *lexical*, *semantic*, *morphological*, *syntactic*, *graphical*, *genre* and *numerical* features). Finally, the third level in the hierarchy refers to specific signals; for example, *reference type* has four specific signals: *personal*, *demonstrative*, *comparative* and *propositional reference*. The hierarchical organization of the signalling taxonomy is provided in Figure 1. Note that subcategories are only illustrative, not exhaustive.

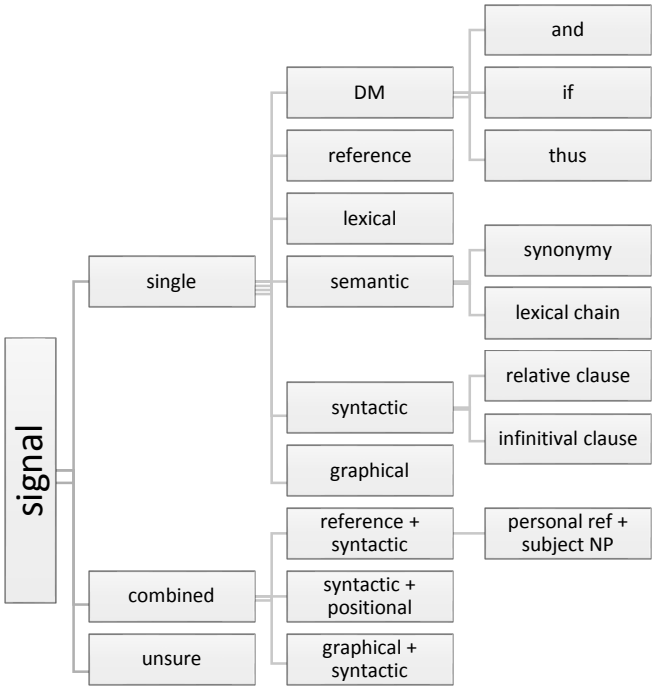

Figure 1: Hierarchical taxonomy of signals

A *single* signal is made of one (and only one) feature used to indicate a particular relation. Consider the following examples from the RST-DT[3]. In Example (11), the DM *although*, which is a single signal, is used to indicate the *Antithesis* relation.

(11) [Although Larsen & Toubro hadn't raised money from the public in 38 years,]S [its new owners frequently raise funds on the local market.]N – Antithesis (wsj_629: 142/143)

In Example (12), the *Contingency* relation is indicated by a lexical signal, the indicative word *contingent*, which represents a single signalling feature.

(12) [Iran's President Rafsanjani offered to help gain freedom for Western hostages in Lebanon,]N [but said the assistance was contingent on U.S. aid in resolving the cases of three Iranians kidnapped in Lebanon in 1982 or the release of frozen Iranian assets.]S – Contingency (wsj_1353: 77/78-82)

The *Purpose* relation in Example (13) is signalled by a syntactic feature, the *infinitival clause* (underlined), which is also a single signal.

(13) [To encourage more competition among exporting countries,]S [the U.S. is proposing that export subsidies, including tax incentives for exporters, be phased out in five years.]N – Purpose (wsj_1135: 54/55-58)

A combined signal[4], on the other hand, comprises two single signals (or features) which work in combination with each other to indicate a particular relation. Consider the following example from the RST-DT.

(14) [Gerald C. Beddall, 47 years old, was named president of the Clairol division of this pharmaceuticals and health-care company.]N [He succeeds C. Benjamin Brooks Jr.,…]S – Elaboration-additional (wsj_1341: 3/4-8)

In this example, two types of single signals, a *reference* feature and a *syntactic* feature, are operative together in signalling the *Elaboration-additional* relation. The reference feature indicates that the word *He* in the satellite span is a personal pronoun because it refers back to Gerald C. Beddall, an entity mentioned (or introduced) in the nucleus span. Syntactically, the personal pronoun, *He*, is also in the subject position of the sentence the satellite span starts with, representing the topic of the *Elaboration-additional* relation. Therefore, the combined signal, comprising the *reference* and *syntactic* features – in

[3] **Conventions for interpreting examples from the RST-DT:** The text within square brackets denotes a span. Each pair of square brackets is followed by either the uppercase character N, referring to the nucleus span, or the uppercase character S, referring to the satellite span. A pair of two spans (N and S) is respectively followed by a dash and the name of the relation that holds between the spans. The relation name is further followed by parentheses containing the file number (of the source document), and the span numbers (the location of the relation in the document), respectively. In addition, the file number and the span numbers within the parentheses are separated by a colon, and each span number is separated from the other span number by a forward slash. For highlighting a particular signal used, the relevant parts (referring to the relevant textual features) in a span are underlined.

[4] A combined signal is represented within parentheses, including two features conjoined by the '+' symbol. For example, a combined signal, containing feature 1 and feature 2, is represented in the following form: (feature 1 + feature 2).

the form of a *personal reference* plus a *subject NP*, represented as (*personal reference + subject NP*) – functions here as a signal for the *Elaboration-additional* relation.

Finally, *unsure* refers to those cases in which no potential signals were found or were specified.

(15) [This hasn't been Kellogg Co.'s year.]S [The oat-bran craze has cost the world's largest cereal maker market share.]N – Cause (wsj_610: 1/2)

(16) ["This is a democratic process]N [-- you can't slam-dunk anything around here."]N – Consequence (wsj_1963: 33/34)

The detailed taxonomy used in our signalling annotation task is provided in Table 2. For more information about the signal class, signal types and specific signals (with definitions, classifications and examples), see Appendix B.

#	Signal class	Signal type	Specific signal
1	single	discourse marker (DM)	and, but, if, since, then, when, etc.
		reference	personal reference demonstrative reference comparative reference propositional reference
		lexical	indicative word alternate expression
		semantic	synonymy antonymy meronymy repetition indicative word pair lexical chain general word
		morphological	tense
		syntactic	relative clause infinitival clause present participial clause past participial clause imperative clause interrupted matrix clause parallel syntactic construction reported speech subject auxiliary inversion nominal modifier adjectival modifier
		graphical	colon semicolon dash parentheses items in sequence
		genre	inverted pyramid scheme newspaper layout newspaper style attribution newspaper style definition
		numerical	same count

#	Signal class	Signal type	Specific signal
2	combined	(reference + syntactic)	(personal reference + subject NP) (demonstrative reference + subject NP) (comparative reference + subject NP) (propositional reference + subject NP)
		(semantic + syntactic)	(repetition + subject NP) (lexical chain + subject NP) (synonymy + subject NP) (meronymy + subject NP) (general word + subject NP)
		(lexical + syntactic)	(indicative word + present participial clause)
		(syntactic + semantic)	(parallel syntactic construction + lexical chain)
		(syntactic + positional)	(present participial clause + beginning) (past participial clause + beginning)
		(graphical + syntactic)	(comma + present participial clause) (comma + past participial clause)
3	unsure	unsure	unsure

Table 2: Taxonomy of signals used in the annotation of relations in the RST-DT

The difference between combined signals and multiple signals is one of independence of operability. In a combined signal, there are two signals, one of which is an independent signal, while the other one is dependent on the first signal. For example, in a combined signal such as (*personal reference + subject NP*), the feature *personal reference* is the independent signal because it directly (and independently) refers back to the entity introduced in the first span. In contrast, the feature *subject NP* is the dependent signal because it is used to specify additional attributes of the first signal. In this particular case, the syntactic role of the personal reference (i.e., a subject NP) in the second span is specified by the use of the second signal *subject NP*. Multiple signals, on the other hand, function independently of and separately from each other, but they all contribute to signalling the relation. For example, in an *Elaboration* relation with multiple signals, involving a genre feature (e.g., *inverted pyramid scheme*) and a lexical feature (e.g., *indicative word*), the signals do not have any connection, as they refer to two different features which separately signal the relation.

4. Annotation Process

In our signalling annotation, we perform a sequence of three tasks: (i) we examine each relation in the RST-DT; (ii) assuming the relational annotation is correct, we search for signals that indicate that such relation is present; and finally (iii) we add to those relations a new layer of annotation of signalling information.

We annotate all the 385 documents in the RST-DT (divided into 347 training documents and 38 test documents) containing 21,400 relations in total. We use the taxonomy of signals (presented in Table 2) to annotate the signals for those relations in the corpus. In some cases, more than one signal may be present. When confronted with a new instance of a particular type of relation, we consulted our taxonomy, and tried to find the appropriate signal(s) that could best function as the indicator(s) for that relation instance. If our search led us to assigning an appropriate signal (or more than one appropriate signal) to that relation, we declared success in identifying the signal(s) for that relation. If our search did not match any of the signals in the taxonomy, then we examined the context (comprising the spans) to discover any potential new signals. If a new signal was identified, we included it in the appropriate category in our

existing taxonomy. In this way, we proceed through identifying the signals of the relations in the corpus, and, at the same time, keep on updating our taxonomy with new signalling information, if necessary.

In order to facilitate the annotation process, we used UAM CorpusTool (O'Donnell, 2008)[5] which provides annotation of texts at multiple levels defined by the user (document layer, semantic-pragmatic level, syntactic level, etc.). The tool can directly import RST files, and show the discourse structure of a text in the form of RST trees, although it does not support layered annotation on top of RST-level structures. As a solution to this problem, we imported the RST base files (along with all relational information) into UAM CorpusTool after converting them from LISP format to a simple text file format. This allowed us to select individual relations and tag them with relevant signal tags. UAM CorpusTool supports a hierarchically-organized tagging scheme (see Section 3 for more detail on the signal taxonomy), and it also provides multiple annotations for a single segment (in our annotation, a large number of relations are indicated by more than one signal). In addition, the annotated data in UAM CorpusTool is stored in XML.

In the annotation process, we import the RST files (in a text file format, converted from the LISP format) into UAM CorpusTool. The visualization window of UAM CorpusTool shows the existing relational annotations, including the RST-segmented texts and the names of the relations holding between text spans. For tagging a particular relation instance, we select the name of the relation, and then choose from the annotation scheme (the taxonomy of signals already incorporated in the tool) the appropriate set of signalling tags (organized into three levels: *signal class*, *signal type* and *specific signal*) in order to assign signalling information to that relation. If the relation contains more signals, we select the relation again (and again, if necessary) and re-do the above-mentioned steps. A snapshot of the annotation window in UAM CorpusTool is provided in Figure 2.

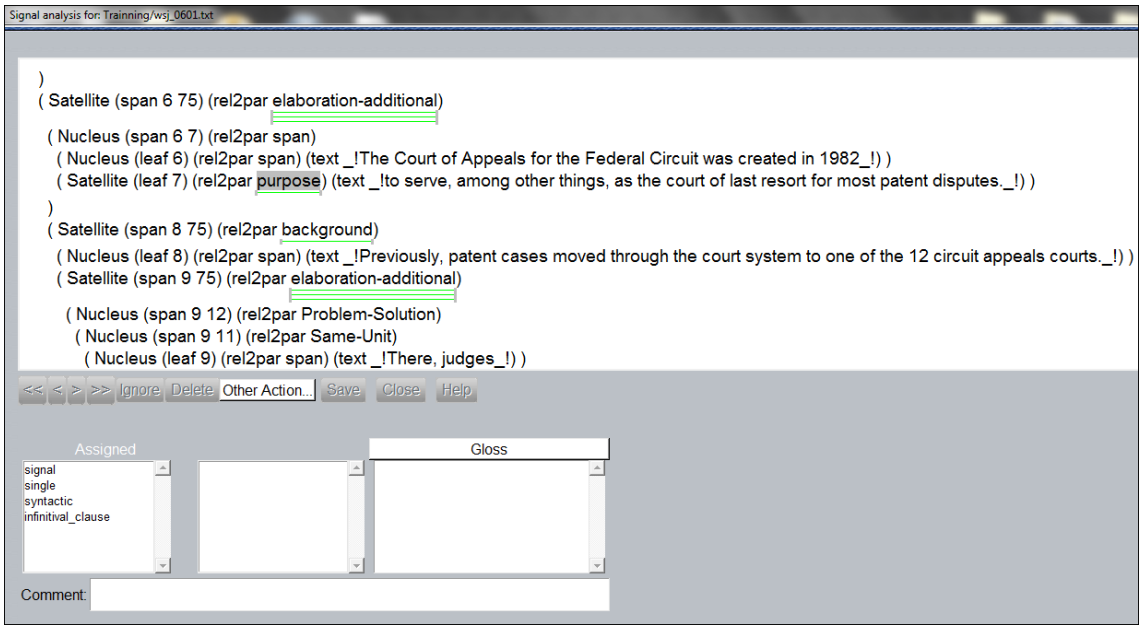

Figure 2: Signalling annotation in UAM CorpusTool

[5] http://www.wagsoft.com/CorpusTool/

We used the 2.8.12 version of UAM CorpusTool to perform our signalling annotation in the RST Signalling Corpus. The annotations are also accessible using the later (and the most recent) versions of UAM CorpusTool, usually by importing the CorpusTool(.ctpr) file in the new version.

Note: Information about the statistical distribution of relations and their signals in the RST Signalling Corpus can be found in Debopam Das' PhD dissertation "Signalling of Coherence Relations in Discourse". The dissertation is available through the SFU library (http://www.lib.sfu.ca/help/publication-types/finding-sfu-theses). The complete distribution of relations and their signals is available from the following URL: http://www.sfu.ca/~mtaboada/research/signalling.html.

5. An Example of Signalling Annotation

We provide the annotation of a short RST file from the RST-DT (file number: wsj_650) with signalling information. The file contains the following text.

(17) Sun Microsystems Inc., a computer maker, announced the effectiveness of its registration statement for $125 million of 6 3/8% convertible subordinated debentures due Oct. 15, 1999.

The company said the debentures are being issued at an issue price of $849 for each $1,000 principal amount and are convertible at any time prior to maturity at a conversion price of $25 a share.

The debentures are available through Goldman, Sachs & Co.

The graphical representation of the RST analysis of the above text using the RSTTool is provided in Figure 3.

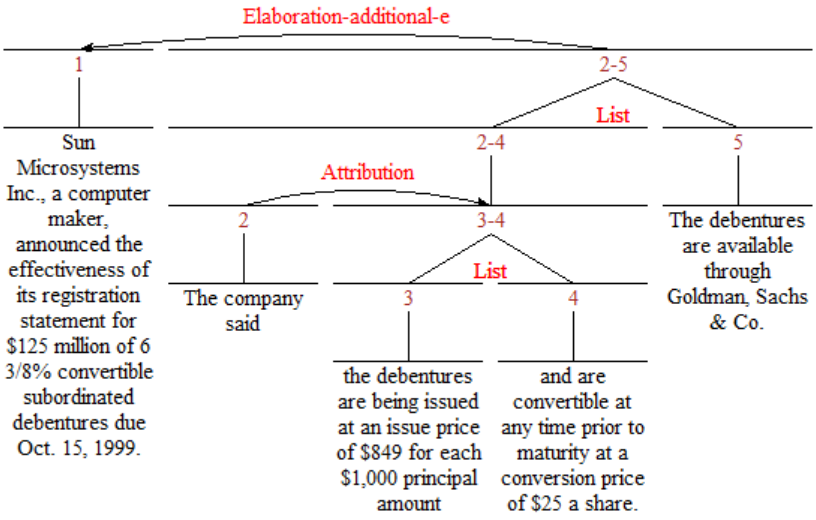

Figure 3: Graphical representation of an RST analysis

The RST analysis shows that the text comprises five spans which are represented in the diagram (in Figure 3) by the numbers, 1, 2, 3, 4 and 5, respectively. In the diagram, the arrowhead points to a span refer to the nuclei, and the arrow points away from another span refer to the satellites. Span 3 (nucleus) and span 4 (nucleus) are in a multinuclear *List* relation, and together they make the combined span 3-4. Span 2 (satellite) is connected to span 3-4 (nucleus) by an *Attribution* relation, and together they make the combined span 2-4. A multinuclear *List* relation holds between spans 2-4 (nucleus) and 5 (nucleus), and together they make the combined span 2-5. Finally, span 2-5 (satellite) is connected to span 1 (nucleus) by an *Elaboration* (more specifically, *Elaboration-addition-e*) relation.

We annotate the relations in the text with appropriate signalling information. A detailed description of our annotation is provided in Table 3.

File	N	S	Relation	Signal type	Specific signal	Explanation: How signalling works
wsj_650	1	2-5	Elaboration-additional	genre	inverted pyramid scheme	In the newspaper genre, the content of the first paragraph (or the first few paragraphs) is elaborated on in the subsequent paragraphs.
				semantic	lexical overlap	The word *debentures* occurs both in the nucleus and satellite.
					lexical chain	Words such as *debentures, issue price, convertible, conversion price* and *share* are in a lexical chain.
				(semantic + syntactic)	(lexical chain + subject NP)	The phrases *Sun Microsystems Inc.* and *the company* in the respective spans are in a lexical chain, and the latter is syntactically used as the subject NP of the sentence the satellite starts with.
	3/4		List	DM	and	The DM *and* functions as a signal for the *List* relation.
	3-4	2	Attribution	syntactic	reported speech	The reporting clause plus the reported clause construction is a signal for the *Attribution* relation.
	2-4/5		List	semantic	lexical chain	The words, *issued, convertible, debentures, available*, in the respective spans are semantically related.

Table 3: Annotation of an RST file with relevant signalling information

According to our annotation, the *Elaboration* relation between spans 1 and 2-5 is indicated by three types of signals, more specifically by two types of *single* signals: *genre* and *semantic* features; and by a *combined* type of signal: (*semantic + syntactic*) feature. First, the text represents the newspaper genre (since it is taken from a Wall Street Journal article). In newspaper texts, the content of the first (or the first few) paragraphs is typically elaborated on in the subsequent paragraphs. A reader, being conscious of the fact that he/she is reading a newspaper text, expects the presence of an *Elaboration* relation between the first paragraph (or the first few paragraphs) and subsequent paragraphs. It is this prior knowledge about the textual organization of the newspaper genre that guides the reader to interpret an *Elaboration* relation between paragraphs in a news text. In this particular example, the entire first paragraph is the nucleus of the *Elaboration* relation, with the two following paragraphs being its satellite. Thus, we postulate that the *Elaboration* relation is conveyed by the *genre* feature (more specifically by a feature which we call *inverted pyramid scheme*). Second, the *Elaboration* relation is also signalled by two *semantic* features: *lexical overlap* and *lexical chain*. The word *debentures* occurs in both the nucleus and satellite spans, indicating the presence of the same topic in both spans, with an elaboration in the second span of some

topic introduced in the first span. Also, words such as *convertible* and *debentures* in the first span and words (or phrases) such as *issue price, convertible, conversion price* and *share* in the second span are semantically related. These words form a lexical chain which is a strong signal for an *Elaboration* relation. Finally, we postulate that a *combined* feature (*semantic + syntactic*), made of two individual features, is operative in signalling the *Elaboration* relation. One can notice that the entity *Sun Microsystems Inc.*, mentioned in the nucleus, is elaborated on in the satellite. The phrase *Sun Microsystems Inc.* is semantically related to the phrase *the company* in the satellite, and hence, they are in a lexical chain. Syntactically, the phrase *the company* is used as the subject NP of the sentence the satellite starts with, representing the topic of the *Elaboration* relation.

The *List* relation between spans 3 and 4 is conveyed in a straightforward (albeit underspecified) way by the use of the DM *and*.

The *Attribution* relation between spans 2 and 3-4 is indicated by a *syntactic* signal, the *reported speech* feature, in which the reporting clause (span 2) functions as the satellite and the reported clause (span 3-4) functions as the nucleus. The key is the subject-verb combination with a reported speech verb (*said*).

Finally, the *List* relation between spans 2-4 and 5 is indicated by a *semantic* feature, *lexical chain*. The words such as *issued* and *convertible* (in the first nucleus) and words *debentures* and *available* (in the second nucleus) are semantically related, indicating (perhaps loosely) a *List* relation between the spans.

6. References

Carlson, L., & Marcu, D. (2001). *Discourse Tagging Manual*.

Carlson, L., Marcu, D., & Okurowski, M. E. (2001). *Building a discourse tagged corpus in the framework of Rhetorical Structure Theory*. Paper presented at the Second SIG dial Workshop on Discourse and Dialogue (SIGdial-2001), Aalborg, Denmark.

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). RST Discourse Treebank, LDC2002T07 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Das, D. (2012). *Investigating the Role of Discourse Markers in Signalling Coherence Relations: A Corpus Study*. Paper presented at the Northwest Linguistics Conference, University of Washington, Seattle.

Das, D., & Taboada, M. (2013). *Explicit and Implicit Coherence Relations: A Corpus Study*. Paper presented at the Canadian Linguistic Association (CLA) Conference, University of Victoria, Canada.

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics, 19*(2), 313-330.

O'Donnell, M. (1997). RSTTool, from http://www.wagsoft.com/RSTTool/

O'Donnell, M. (2008). *The UAM CorpusTool: Software for corpus annotation and exploration.* Paper presented at the XXVI Congreso de AESLA, Almeria, Spain.

Taboada, M., & Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse, 4*(2), 249-281.

7. Appendix A

List of Relations in the RST-DT

The RST Discourse Treebank contains 78 relation types, divided into 53 mononuclear and 25 multinuclear relations. Table 4 provides the complete listing of relations, arranged alphabetically by mononuclear relations (reproduced from Carlson and Marcu (2001: 42-44)). 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 mononuclear and multinuclear relations are shown across a single row. In addition, mononuclear relation names begin with a lowercase letter, and multinuclear relation names begin with an uppercase letter.

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		
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		

Mononuclear (satellite)	Mononuclear (nucleus)	Multinuclear
means		
otherwise		Otherwise
preference		
problem-solution-s	problem-solution-n	Problem-Solution
		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

Table 4: List of relations in the RST Discourse Treebank

8. Appendix B

Signals of Coherence Relations

In this section, the description of the signals used for annotating signals of coherence relations in the RST-DT is provided. The description includes definitions of the signals, their classifications and examples from the corpus.

Conventions for interpreting examples from the RST-DT: The text within square brackets denotes a span. Each pair of square brackets is followed by either the uppercase character N, referring to the nucleus span, or the uppercase character S, referring to the satellite span. A pair of two spans (N and S) is respectively followed by a dash and the name of the relation that holds between the spans. The relation name is further followed by parentheses containing the file number (of the source document), and the span numbers (the location of the relation in the document), respectively. In addition, the file number and the span numbers within the parentheses are separated by a colon, and each span number is separated from the other span number by a forward slash. For highlighting a particular signal used, the relevant parts (referring to the relevant textual features) in a span are underlined.

1. Single Signals

A single signal is made of one (and only one) feature used to indicate a particular relation. The class of single signals comprises nine types of signals.

1.1. Discourse marker

Definition: Discourse Markers (DMs) are lexical expressions which are primarily drawn from syntactic categories, such as conjunctions, adverbials and prepositional phrases. DMs connect discourse segments, and signal a coherence relation between those segments. E.g.,

(18) [Although Larsen & Toubro hadn't raised money from the public in 38 years,]S [its new owners frequently raise funds on the local market.]N – Antithesis (wsj_629: 142/143)

DMs generally occur at the beginning of a span. However, they can also occur in the middle of a span. E.g.,

(19) [Lawmakers often are reluctant to embarrass colleagues, even those of opposing political parties.]N [In the recent Housing and Urban Development Department scandal, for example, Rep. Thomas Lantos, the California Democrat who led the hearings, tiptoed through embarrassing disclosures about HUD grants secured by Sen. Alfonse D'Amato, a New York Republican.]S – Example (wsj_1366: 31/32-35)

(20) [Mr. Lee, president of Luzon Petrochemical Corp., said the contract was signed Wednesday in Tokyo with USI Far East officials.]N [Contract details, however, haven't been made public.]S – Elaboration-additional (wsj_606: 10-11/12)

DMs can also occur at the end of a span (although very rarely). E.g.,

(21) [In this connection, it is important to note that several members of New York's sitting City Council represent heterogeneous districts that bring together sizable black, Hispanic, and non-Hispanic white populations]N [-- Carolyn Maloney's 8th district in northern Manhattan and the south Bronx and Susan Alter's 25th district in Brooklyn, for example.]S – Example (wsj_1137: 78-79/80)

DMs can occur discontinuously as a pair in which one part presupposes the presence of other, while both parts, in combination, signal a single relation. E.g.,

(22) [interest costs would either be paid by the student]N [or added to the loan balance.]N – Otherwise (wsj_1131: 91/92)

(23) [If the economy slips into a recession,]S [then this isn't a level that's going to hold.]N – Condition (wsj_681: 156/157)

(24) [Not only are there camera operators on all sides,]N [but the proceedings are shown on monitors throughout the theater.]N – List (wsj_1984: 15/16)

Sometimes, two DMs are conjoined with each other, and can be used as a single DM. E.g.,

(25) [When and if the trust runs out of cash -- which seems increasingly likely --]S [it will need to convert its Manville stock to cash.]N – Condition (wsj_1328: 16-17/18)

Sometimes, two DMs can be used separately to indicate a single relation, as in the case of multiple signals. E.g.,

(26) [that would allow them to acknowledge that Sverdlovsk violated the 1972 agreement]N [or, alternatively, that would give U.S. specialists reasonable confidence that this was a wholly civilian accident.]N – Disjunction (wsj_1143: 78-79/80-81)

(27) [it expects to shed its remaining mortgage loan origination operations outside its principal markets in New Jersey and Florida]N [and, as a result, is taking a charge for discontinued operations.]S – Cause (wsj_2359: 25/26)

1.2. Reference features

Reference features used in our annotation are based on the concept of reference (under grammatical cohesion) as proposed in Halliday and Hasan (1976). Reference items are represented by pronouns and other referential expressions. We used four types of reference: *personal reference*, *demonstrative reference*, *comparative reference* and *propositional reference* (used in the sense of *extended reference* and *text reference* in Halliday and Hasan (1976)).

1.2.1. Personal Reference

Definition: The personal reference feature refers to pronouns (such as *I, they and him*), possessive determiners (such as *my, your and her*) and possessive pronouns (such as *mine and yours*) which are present in one span, and refer to an object or entity (or a pronoun) mentioned in the other span. E.g.,

(28) [Michael A. Miles, chief executive officer of Philip Morris Cos. ' Kraft General Foods unit, bought 6,000 shares of the company on Sept. 22 for $157 each.]N [The $942,000 purchase raised his holdings to 74,000 shares.]S – Elaboration-additional (wsj_1157: 68/69)

(29) [Many adjusters are authorized to write checks for amounts up to $100,000 on the spot.]N [They don't flinch at writing them.]S – Elaboration-additional (File 3: 27/28)

(30) [Whatever the difficulties, Mr. Gorbachev remains committed to increasing foreign trade.]N [For political as well as economic reasons, U.S. companies are at the top of his priorities -- a point he underscored by spending two hours walking around the U.S. trade show last week.]S – Elaboration-additional (wsj_1368: 49/50-52)

1.2.2. Demonstrative Reference

Definition: The demonstrative reference feature refers to four demonstrative determiners: *this, that, these* and *those*, four demonstrative pronouns: *this, that, these* and *those*, and four adverbs: *here, there, now* and *then*, which are present in one span and refer to an object or entity mentioned in the other span. E.g.,

(31) [Six top executives at the New York-based company sold shares in August and September.]N [Four of those insiders sold more than half their holdings.]S – Elaboration-general-specific (wsj_1157: 2/3)

(32) [(ABC stops short of using an "applause" sign and a comic to warm up the audience.]N [The stars do that themselves.)] – Elaboration-additional (wsj_633: 66-67/68)

(33) [… Adjusters must count the number of bathrooms, balconies, fireplaces, chimneys, microwaves and dishwashers.]N [But they must also assign a price to each of these items as well as to floors, wallcoverings, roofing and siding, to come up with a total value for a house…]S – Elaboration-additional (wsj_File3: 110-111/112-115)

(34) [The argument turns on the discovery in 1909 of an amazing fossil quarry high in the Canadian Rockies called the Burgess Shale.]N [Here, in an area smaller than a city block, lay buried traces of countless weird creatures that had frolicked more than 500 million years ago…]S – Elaboration-additional (wsj_1158: 26-27/28-31)

1.2.3. Comparative Reference

Definition: The comparative reference feature refers to those reference items (words such as *equal, identical, similar, differently, more, less, better* and *worse*) which are present in one span and refer to an object or entity in the other span by means of identity or similarity. E.g.,

(35) [We are working significantly longer and harder]N [than has been the case in the past]S – Comparison (wsj_604: 81/82)

(36) [Texas and California are traditionally powerful within the conference,]N [but equally striking is the dominance of Alaska, Pennsylvania and West Virginia because of their power elsewhere in the appropriations process.]N – Comparison (wsj_1147: 84/85-86)

(37) [In a great restaurant, don't deprive yourself.]N [The other meals don't matter.]N – Contrast (wsj_1367: 148/149)

1.2.4. Propositional Reference

Definition: The propositional reference feature, usually represented by pronouns: *it, this* and *that*, in one span, refers to a proposition (a process, phenomenon or fact, and NOT an object or entity) in the other span. E.g.,

(38) [They've been looking to get their costs down,]N [and this is a fairly sensible way to do it]S – Elaboration-additional (wsj_2394: 46/47)

(39) ["They've been looking to get their costs down,]N [and this is a fairly sensible way to do it,"]S – Elaboration-additional (wsj_2394: 46/47)

(40) [An official with lead underwriter First Boston said orders for the San Antonio bonds were "on the slow side."]N [He attributed that to the issue's aggressive pricing and large size, as well as the general lethargy in the municipal marketplace…]S – Elaboration-additional (wsj_1322: 136-137/138-144)

1.3. Lexical features

Lexical features include the use of indicative words and phrases, such as individual words that indicate a relation, for example, the verbs *concede* and *cause* for *Concession* and *Cause* respectively. The difference between DMs and lexical features is that DMs are used as the linking elements between discourse segments, but items representing lexical features do not connect text spans.

1.3.1. Indicative Word

Definition: The indicative word feature refers to a word or phrase which signals a relation. E.g.,

(41) [Sales in the first half came to 159.92 billion yen,]N [compared with 104.79 billion yen in the four-month period.]N – Comparison (wsj_643: 11/12)

(42) [Iran's President Rafsanjani offered to help gain freedom for Western hostages in Lebanon,]N [but said the assistance was contingent on U.S. aid in resolving the cases of three Iranians kidnapped in Lebanon in 1982 or the release of frozen Iranian assets.]S – Contingency (wsj_1353: 77/78-82)

(43) [Mr. Palmero recommends Temple-Inland,]N [explaining that it is "virtually the sole major paper company not undergoing a major capacity expansion," and thus should be able to lower long-term debt substantially next year.]S – Explanation-argumentative (wsj_666: 72/73-76)

1.3.2. Alternate Expression

Definition: The alternate expression feature refers to a short tensed clause which functions as the signal of a relation. E.g.,

(44) ["much of the increase in debt in recent years is due to increasing credit use by higher-income families,"]N [that is, "those probably best able to handle it."] – Elaboration-additional (wsj_1389: 56/57)

(45) [Production of full-sized vans will be consolidated into a single plant in Flint, Mich.]N [That means two plants -- one in Scarborough, Ontario, and the other in Lordstown, Ohio -- probably will be shut down after the end of 1991…]S – Interpretation (wsj_2338: 45: 46-53)

(46) [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.]N

[The result is that instead of the Pentagon's proposed split of 60-40 between domestic and foreign bases, the reduced funding is distributed by a ratio of approximately 70-30.]N – Cause-result (wsj_686: 10-11/12)

1.4. Semantic features

Unlike most other single signals, a semantic feature has two components (words or phrases), each belonging to one of the spans. The components are in a semantic relationship with each other, such as synonymy, antonymy and lexical chain. Unlike reference features which are exclusively represented by pronouns and other referential expressions, the *semantic* features are represented by devices of lexical cohesion (Halliday & Hasan, 1976), and not by pronouns and referential expressions.

1.4.1. Synonymy

Definition: Words or phrases in respective spans are in a synonymy relationship, or a proper noun or a name in one span is abbreviated or mentioned as an acronym (referring to the same object or entity) in the other span. E.g.,

(47) [Tandy's stock fell… Net for the quarter was $62.8 million, or 73 cents a share, down from $64.9 million, or 72 cents a share, a year earlier.]N [The company said earnings would have increased if it hadn't been...]S – Elaboration-additional (wsj_1374: 9-12/13-17)

(48) [Finnair, Finland's state-owned airline, joined the wave of global airline alliances and signed a wide-ranging cooperation agreement with archrival Scandinavian Airlines System.]N

[Under the accord, Finnair agreed to coordinate flights, marketing and other functions with SAS, the 50%-state-owned airline of Denmark, Norway and Sweden.]S – Elaboration-additional (wsj_631: 1-2/3)

1.4.2. Antonymy

Definition: Words or phrases in respective spans are in an antonymy relationship. E.g.,

(49) [In 1988, Kidder eked out a $46 million profit, mainly because of severe cost cutting.]N [Its 1,400-member brokerage operation reported an estimated $5 million loss last year,…]N – Contrast (wsj_604: 31-32/33-40)

(50) [While oil prices have been better than expected,]S [natural gas prices have been worse.]N – Antithesis (wsj_2325: 61-62/63)

(51) [… higher$_i$ bidding narrows$_i$ the investor's return]N [while lower$_i$ bidding widens$_i$ it.]N – Contrast (wsj_1322: 82/83)

1.4.3. Meronymy

Definition: Words or phrases in respective spans are in a meronymy relationship. In other words, a set of objects or entities is introduced in one span, and a member object or entity from that set is mentioned in the other span. E.g.,

(52) [Predicting the financial results of computer firms has been a tough job lately.]N

[Take Microsoft Corp., the largest maker of personal computer software and generally considered an industry bellwether…]S – Example (wsj_2365: 1/2-22)

(53) [… it gave ground to Mr. Inouye on a number of projects,]N [ranging from a $11 million parking garage here, to a land transfer in Hawaii, to a provision to assist the Makwah Indian Tribe in Washington state.]S – Example (wsj_1147: 98/99-103)

1.4.4. Repetition

Definition: An entity is introduced in one span, and the entity (or its name) is repeated in the other span. E.g.,

(54) ["They are not a happy group of people at Battle Creek right now."]N [Kellogg is based in Battle Creek, Mich., a city that calls itself the breakfast capital of the world.] – Elaboration-additional (wsj_610: 29/30-31)

(55) [Industry estimates put Avis's$_i$ annual cost$_j$ of all five programs at between $8 million and $14 million.]N [A spokesman for Avis$_i$ wouldn't specify the costs$_j$ but said the three airlines being dropped account for "far less than half" of the total.]S – Elaboration-additional (wsj_2394: 8/9-13)

1.4.5. Indicative Word Pair

Definition: Words or phrases in the respective spans form a word (or phrasal) pair as they are very closely related by their semantic content. E.g.,

(56) [Under one count, Gulf Power would plead guilty to conspiring to violate the Utility Holding Company Act.]N [Under the second count, the company would plead guilty to conspiring to evade taxes.]N – List (wsj_619: 16/17)

(57) [Asked about the speculation that Mr. Louis-Dreyfus has been hired to pave the way for a buy-out by the brothers,]S [the executive replied, "That isn't the reason Dreyfus has been brought in. He was brought in to turn around the company."]N – Question-Answer (wsj_2331: 44-46/47-51)

(58) [… Mr. Lawson resigned from his six-year post because of a policy squabble with other cabinet members.]N

[He was succeeded by John Major, who Friday expressed a desire for a firm pound and supported the relatively high British interest rates…]N – Sequence (wsj_693: 51-52/53-61)

1.4.6. Lexical Chain

Definition: Words or phrases in the respective spans are identical or semantically related. Unlike the repetition feature, words or phrases in lexical chains do not refer to object or entity, but they belong to the class of indefinite or common nouns and also other syntactic categories, such as adjectives, verbs and adverbs. Lexical chain differs from synonymy, antonymy, meronymy and indicative word pair in a significant way. While in the latter categories, the semantic relation between the words or phrases is very strong and can be specified, words or phrases present in a lexical chain are related to each other by a relatively weak semantic connection. E.g.,

(59) [Insiders have been selling$_i$ shares$_j$ in Dun & Bradstreet Corp., the huge credit$_l$-information concern$_k$.]N

[Six top executives at the New York-based company$_k$ sold$_i$ shares$_i$ in August and September. Four of those insiders sold$_i$ more than half their holdings$_i$.]S – Elaboration-general-specific (wsj_1157: 1/2-3)

(60) [Personal-computer makers will continue to eat away at the business of more traditional computer firms.]N [Ever-more powerful desk-top computers, designed with one or more microprocessors as their "brains," are expected to increasingly take on functions carried out by more expensive minicomputers and mainframes.]S – Elaboration-additional (wsj_2365: 109/110-113)

Sometimes, the connection between words or phrases present in a lexical chain is more of a suggestive nature specifying a cause-result, explanatory or contrast relationship. E.g.,

(61) [Since commercial airline flights were disrupted,]S [the company chartered three planes to fly these executives back to the West Coast and bring along portable computers, cellular phones and some claims adjusters.]N – Reason (wsj_File 3: 84/85-87)

(62) [Robert F. Singleton, Knight-Ridder's chief financial officer, said the company was "pleased" with its overall performance, despite only single-digit growth in newspaper revenue.]N [That division's revenue rose 2.3% to $472.5 million from $461.9 million in the year-ago period. Gains in advertising revenue, however, resulted in operating profit of $78.4 million -- up 20% from $65.6 million.]S – Explanation-argumentative (wsj_1182: 9/10-12)

(63) [… he accepted the resignation of Thomas Wilson, vice president of corporate sales,]N [… his marketing responsibilities have been reassigned]N – Sequence (wsj_2342: 19/20)

(64) [While the earnings picture confuses,]S [observers say the major forces expected to shape the industry in the coming year are clearer.]N – Antithesis (wsj_2365: 47/48-51)

1.4.7. General Word

Definition: Words such as *thing*, *matter* and *issue* which are present in one span, and refer to an object, entity, fact or proposition in the other span in a more general way. E.g.,

(65) ["You have to count everything." Adjusters must count the number of bathrooms, balconies, fireplaces, chimneys, microwaves and dishwashers.]N [But they must also assign a price to each of these items as well as to floors, wallcoverings, roofing and siding, to come up with a total value for a house. To do that, they must think in terms of sheetrock by the square foot, carpeting by the square yard, wallpaper by the roll, molding by the linear foot.]S – Elaboration-additional (wsj_File3: 110-111/112-115)

(66) [Italian President Francesco Cossiga promised a quick investigation into whether Olivetti broke Cocom rules.]N [President Bush called his attention to the matter during the Italian leader's visit here last week.]S – Elaboration-additional (wsj_2326: 28-29/30)

1.5. Morphological features

1.5.1. Tense

Definition: The tense feature refers to a change of tense, aspect or mood between the relevant clauses or sentences in the respective spans. E.g.,

(67) [In June, the company agreed to settle for $18 million several lawsuits related to its sales practices, without admitting or denying the charges.]N [An investigation by U.S. Postal inspectors is continuing.]N – Sequence (wsj_1157: 21-23/24)

(68) [Neither suit lists specific dollar claims,]N [largely because damage assessment hasn't yet been completed.]S – Circumstance (wsj_1347: 10/11)

(69) [… the opposition parties are so often opposed to whatever LDP does]N [that it would be a waste of time.]S – Consequence (wsj_1120: 70/71)

1.6. Syntactic features

1.6.1. Relative Clause

Definition: One span, functioning as the satellite, is a relative clause modifying an object or entity (or a proposition in a few instances) present in the other span or nucleus. E.g.,

(70) [One of Dun & Bradstreet's chief businesses is compiling reports]N [that rate the credit-worthiness of millions of American companies.]S – Elaboration-object-attribute (wsj_1157: 12/13)

(71) [The Tokyo-based company had net of 3.73 billion yen in the previous reporting period,]N [which was the four months ended March 31.]S – Elaboration-additional (wsj_643: 8/9-10)

1.6.2. Infinitival Clause

Definition: One span, functioning as the satellite, is an infinitival clause embedded under the main clause or nucleus. E.g.,

(72) [… this tactic was designed]N [to soften the blow of declining stock prices and generate an offsetting profit by selling waves of S&P futures contracts.] – Purpose (wsj_2381: 86/87-89)

(73) [To encourage more competition among exporting countries,]S [the U.S. is proposing that export subsidies, including tax incentives for exporters, be phased out in five years.]N – Purpose (wsj_1135: 54/55-58)

1.6.3. Present Participial Clause

Definition: One span, functioning as the satellite, is a present participial clause embedded under the main clause or nucleus. E.g.,

(74) [Wyse has done well]N [establishing a distribution business,]S – Manner (wsj_2365: 99/100)

(75) [NASA pronounced the space shuttle Atlantis ready for launch tomorrow]N [following a five-day postponement of the flight because of a faulty engine computer.]S – Circumstance (wsj_2356: 48/49-50)

1.6.4. Past Participial Clause

Definition: One span, functioning as the satellite, is a past participial clause embedded under the main clause or nucleus. E.g.,

(76) [The offer would give the transaction an indicated value of $189 million,]N [based on the 18.9 million shares the group doesn't already own.]S – Circumstance (wsj_697:4/5-6)

(77) [wedged between shifting dunes and pounding waves at the world's most inhospitable diamond dig,]S [lies the earth's most precious jewel box.]N – Elaboration (wsj_1121: 21/22)

1.6.5. Imperative Clause

Definition: One span, functioning as the satellite, is an imperative clause. E.g.,

(78) [Predicting the financial results of computer firms has been a tough job lately.]N
[Take Microsoft Corp., the largest maker of personal computer software and generally considered an industry bellwether…]S – Example (wsj_2365:1/2-22)

(79) ["Now you go to districts,]S [you're likely to get candidates whose views are more extreme, white and black, on racial issues."]N – Contingency (wsj_1137: 114/115-116)

1.6.6. Interrupted Matrix Clause

Definition: The nucleus span is a sentence which is interrupted by the insertion of a clause or phrase functioning as the satellite span. E.g.,

(80) [At a recent meeting of manufacturing executives, "everybody]N_i [I talked with]S [was very positive,"]N_i – Same-Unit (wsj_628: 25/27)

(81) [Litigation,]N [if not settled out of court,]S [could drag on for years.]N – Same-Unit (wsj_1347: 14/16)

1.6.7. Parallel Syntactic Construction

Definition: The spans (clausal segments) or part of the spans (phrasal segments) are parallel to each other in syntactic construction. E.g.,

(82) [that only a black politician can speak for a black person,]N [and that only a white politician can govern on behalf of a white one.]N – List (wsj_1137:43/44)

(83) [Time may seek to break up the transaction after it is consummated,]N [or may seek constraints that would prevent...]N – Disjunction (wsj_1190:28-29/30-31)

Sometimes, similar syntactic constructions such as a pair of reported speeches or imperative clauses or interrogative clauses are also considered to form a parallel syntactic construction. E.g.,

(84) [Gaylord Container said analysts are skeptical of it because it's carrying a lot of debt.]N [Champion International said, "We've gotten our costs down and we're better positioned for any cyclical downturn than we've ever been."]N – List (wsj_666: 46-48/49-52)

(85) [Do you really need this much money to put up these investments?]N [Have you told investors what is happening in your sector?]N – List (wsj_629: 130-131/132-133)

1.6.8. Reported Speech

Definition: The satellite span is the reporting speech and the nucleus span is the reported speech. E.g.,

(86) [Legal strategists say]S [that damage claims against the oil giant and others could well exceed $1 billion.]N – Attribution (wsj_1347: 12/13)

(87) [Machine tool executives are hopeful, however,]S [that recent developments in Eastern Europe will expand markets for U.S.-made machine tools in that region.]N –Attribution (wsj_628: 38/39)

(88) [September orders for machine tools rebounded from the summer doldrums, but remained 7.7% below year-earlier levels,]N [according to figures from NMTBA -- the Association for Manufacturing Technology.]S – Attribution (wsj_628: 4-5/6-7)

1.6.9. Subject Auxiliary Inversion

Definition: The position of the subject and auxiliary verb in a subordinate clause (functioning as the satellite) is interchanged. E.g.,

(89) [Should the courts uphold the validity of this type of defense,]S [ASKO will then ask the court to overturn such a vote-diluting maneuver recently deployed by Koninklijke Ahold NV.]N – Condition (wsj_2383: 11/12-13)

(90) [Had he been a little less gung-ho,]S ["I'd have gotten the thing on the ground and headed for the nearest bar," Mr. Brown says.]N – Condition (wsj_1394: 15/16-18)

1.6.10. Nominal Modifier

Definition: The satellite span is a reduced relative clause or a non-finite clause functioning as the modifier of an object or entity present in the main clause or nucleus. E.g.,

(91) [state officials interfered with the oil company's initial efforts]N [to treat last spring's giant oil spill.]S – Elaboration-object-attribute (wsj_1347: 3/4)

(92) [The action is a counterclaim to a suit]N [filed by Alaska in August against Exxon and six other oil companies.]S – Elaboration-object-attribute (wsj_1347: 5/6)

1.6.11. Adjectival Modifier

Definition: The satellite span is a non-finite clause functioning as the modifier of an adjective present in the main clause or nucleus. E.g.,

(93) [it is prudent]N [to plan for next year on the assumption that revenue again will be flat.]S – Elaboration-additional (wsj_1155: 7/8-9)

(94) [Conviction on any single impeachment article was enough]N [to remove Judge Hastings from office.]S – Elaboration-object-attribute (wsj_1396: 12/13)

1.7. Graphical features

1.7.1. Colon

Definition: The first span ends with a colon followed by the second span. E.g.,

(95) [The market has taken two views:]N [that the labor situation will get settled in the short term and that things look very rosy for Boeing in the long term,]S – Elaboration-set-member (wsj_2308: 78/79-80)

1.7.2. Semicolon

Definition: The first span ends with a colon followed by the second span. E.g.,

(96) [The standard of living has increased steadily over the past 40 years;]N [more than 90% of the people consider themselves middle class.]N – List (wsj_1120: 20/21)

1.7.3. Dash

Definition: The first span ends with a dash followed by the second span, or one of the spans is within dashes. E.g.,

(97) [For political as well as economic reasons, U.S. companies are at the top of his priorities]N [-- a point he underscored by spending two hours walking around the U.S. trade show last week.]S – Elaboration-additional (wsj_1368: 50/51-52)

1.7.4. Parentheses

Definition: The satellite span is inside parentheses. E.g.,

(98) [The expected amount is said to be 700 billion yen]N [($4.93 billion)]S – Restatement (wsj_1187: 53/54)

1.7.5. Items in Sequence

Definition: The nuclei in a multinuclear relation are presented as a numbered list or as items occurring in a sequential order. E.g.,

(99) [B & H Crude Carriers Ltd. -- Four million common shares, via Salomon Brothers Inc.]N

[Chemical Banking Corp. -- 14 million common shares, via Goldman, Sachs & Co.]N

[Chemex Pharmaceuticals Inc. -- 1.2 million units consisting of two shares of common stock and one common warrant, via PaineWebber Inc.]N – List (wsj_678: 7-8/9-10/11-14)

1.8. Genre features

1.8.1. Inverted Pyramid Scheme

Definition: The content of the first paragraph (or the first few paragraphs) is elaborated on in the subsequent paragraphs. Typically in a newspaper report or news article, the most important information (or the topics) is presented as a summary in the beginning of the text or in the first paragraph (or in the first few paragraphs), and the more detail is provided about those information (or those topics) in the paragraphs that follow. E.g.,

(100) [Sun Microsystems Inc., a computer maker, announced the effectiveness of its registration statement for $125 million of 6 3/8% convertible subordinated debentures due Oct. 15, 1999.]N

[The company said the debentures are being issued at an issue price of $849 for each $1,000 principal amount and are convertible at any time prior to maturity at a conversion price of $25 a share.

The debentures are available through Goldman, Sachs & Co.]S – Elaboration-additional (wsj_650: 1/2-5)

1.8.2. Newspaper Layout

Definition: Visual features that helps understanding the organization of a newspaper text (e.g., heading, date and place, body of text, information about the author). E.g.,

(101) [Companies listed below reported quarterly profit substantially different from the average of analysts' estimates. The companies are followed by at least three analysts, and had a minimum five-cent change in actual earnings per share. Estimated and actual results involving losses are omitted…]N

[Source: Zacks Investment Research]N – TextualOrganization (wsj_696: 1-11/12-13)

(102) [Monday, October 23, 1989]N

[The key U.S. and foreign annual interest rates below are a guide to general levels but don't always represent actual transactions…]N – TextualOrganization (wsj_1339: 1/2-70)

1.8.3. Newspaper Style Attribution

Definition: Features characteristic of the newspaper genre, indicative of *Attribution* relations. E.g.,

(103) ["Dividend News: Payout Stalled at Quantum Chemical Corp. --- Firm Posts Quarterly Loss, Plans a Stock Dividend to Take Place of Cash"]N [-- WSJ Oct. 27, 1989)]S – Attribution (wsj_614: 4-8/9)

(104) [Debate on IRAs Centers on Whether Tax Break Should Be Immediate or Put Off Till Retirement"]N [-- WSJ Oct. 27, 1989)]S – Attribution (wsj_605: 6/7)

1.8.4. Newspaper Style Definition

Definition: Features characteristic of the newspaper genre, indicative of *Definition* relations. E.g.,

(105) [PRIME RATE: 10 1/2%.]N [The base rate on corporate loans at large U.S. money center commercial banks.]S – Definition (wsj_1118: 4-5/6)

(106) [MERRILL LYNCH READY ASSETS TRUST: 8.59%.]N [Annualized average rate of return after expenses for the past 30 days; not a forecast of future returns.]S – Definition (wsj_1118: 67-68/69)

1.9. Numerical features

1.9.1. Same Count

Definition: The number of certain objects or entities represented by a word (e.g., *five, two*) in one span is equal to the numerical count of those objects or entities present in the other span. E.g.,

(107) [The investor group includes Restaurant Investment Partnership, a California general partnership, and three Rally's directors:]N [Mr. Sugarman, James M. Trotter III and William E. Trotter II.]S – Elaboration-set-member (wsj_695: 11/12)

(108) [That means two plants]N [-- one in Scarborough, Ontario, and the other in Lordstown, Ohio --]S – Elaboration-general-specific (wsj_2338: 46/47)

2. Combined Signals

A combined signal comprises two single (other) signals (or features) which work in combination with each other to indicate a particular relation. The class of combined signals includes six types of signals.

2.1. (reference + syntactic) features

2.1.1. (personal reference + subject NP)

Definition: An object or entity (or a pronoun) is mentioned in the first (also nucleus) span, and a personal pronoun (*I, she, they*) referring to the same object or entity (or that previously mentioned pronoun) is used as (i) the subject NP of the sentence the satellite span starts with, or (ii) the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with.

Also, a personal pronoun (*I, she, they*) is used in the first (also nucleus) span, and an object or entity referring to the same pronoun is used as (i) the subject NP of the sentence the satellite span starts with, or (ii) the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(109) [John C. Holt, an executive vice president and Dun & Bradstreet director, sold 10,000 shares on Aug. 31 for $588,800, filings show.]N [He retains 9,232 shares.]S – Elaboration-additional (wsj_1157: 29-30/31)

(110) [He added that the company expects "strong" operating profit for the year, "but at a level significantly lower than last year."]N [He said 1989's net income could be 11% to 13% of revenue, ...] S – Elaboration-additional (wsj_1155: 35-36/37-42)

(111) ["They are not a happy group of people at Battle Creek right now."]N [Kellogg is based in Battle Creek, Mich., a city that calls itself the breakfast capital of the world.]S – Elaboration-additional (wsj_610: 29/30-31)

2.1.2. (demonstrative reference + subject NP)

Definition: An object or entity (or demonstrative pronoun) is mentioned in the first (also nucleus) span, and a demonstrative pronoun (*this, that, those*) referring to the same object or entity (or that previously mentioned pronoun) is used as (i) the subject NP of the sentence the satellite span

starts with, or (ii) the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(112) ["We're talking policy limits."]N [In this case, that's about $250,000.]S – Elaboration-general specific (wsj_File3: 189/190)

(113) [The issue includes $100 million of insured senior lien bonds.]N [These consist of current interest bonds due 1990-2002, 2010 and 2015, and capital appreciation bonds due 2003 and 2004, …]S – Elaboration-general-specific (wsj_1161: 69/70-73)

(114) [this is not bad news;]N [this is a blip,"]S – Elaboration-additional (wsj_2358: 43/44)

2.1.3. (comparative reference + subject NP)

Definition: An object or entity is introduced in the first (also nucleus) span, and a comparative referential item (e.g., *other, another*) referring to the same object or entity is used as (i) the subject NP of the sentence the satellite span starts with, or (ii) the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(115) [Hundreds of East Germans flocked to Bonn's Embassy in Warsaw, bringing to more than 1,200 the number of emigres expected to flee to the West beginning today.]N [More than 2,100 others escaped to West Germany through Hungary over the Weekend.]S – Elaboration-additional (wsj_2356: 67-70/71)

2.1.4. (propositional reference + subject NP)

Definition: A fact, process or proposition in the first span (also nucleus) is referred to by the pronouns *it, this* or *that*, and the pronoun also occurs as (i) the subject NP of the sentence the satellite span starts with, or (ii) the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(116) [Introducing pool, argued Councilwoman Riley Reinker, would be "dangerous.]N [It would open a can of worms."…]S – Elaboration-additional (wsj_2367: 32-34/35-36)

(117) [the network now needs to "broaden the horizons of nonfiction television,]N [and that includes some experimentation."]S – Elaboration-additional (wsj_633: 47:48)

(118) [… Manhattan retail rents have dropped 10% to 15% in the past six months alone, experts say.]N [That follows a more subtle decline in the prior six months, after Manhattan rents had run up rapidly since 1986.]S – Elaboration-additional (wsj_2346: 27-28/29-30)

2.2. (semantic + syntactic) features

2.2.1. (repetition + subject NP)

Definition: An object or entity in the first (also nucleus) span is repeated, and it occurs as (i) the head of the subject NP of the sentence the satellite span starts with, or (ii) the head of the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(119) [Machine tool shipments last month were $281.2 million, a 24% rise from a year earlier and a 25% increase from August.]N [Shipments have run well ahead of 1988 all year, as machine tool builders produce against relatively good backlogs…]S – Elaboration-additional (wsj_628: 62/63-71)

(120) [Craig Tillery, an Alaska assistant attorney general, said in an interview last night that Exxon's accusations "are not new.]N [Exxon has made them before, at which point the state demonstrated they were untrue.]S – Elaboration-additional (wsj_1347: 25-26/27-29)

(121) [The company is beginning to ship a new software program that's being heralded as a boon for owners of low-end printers sold by Apple…]N [John Warnock, Adobe's chief executive officer, said the Mountain View, Calif., company has been receiving 1,000 calls a day about the product...]S – Elaboration-additional (wsj_2365: 63-66/67-69)

2.2.2. (lexical chain + subject NP)

Definition: A word (or phrase) in the satellite (also second) span is either identical or semantically related to a certain word(s) present in the nucleus (also first) span, and that word (or phrase) in the satellite span is used as (i) the head of the subject NP of the sentence the satellite span starts with, or (ii) the head of the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(122) [Michael A. Miles, chief executive officer of Philip Morris Cos. ' Kraft General Foods unit, bought 6,000 shares of the company on Sept. 22 for $157 each.]N [The $942,000 purchase raised his holdings to 74,000 shares.]S – Elaboration-additional (wsj_1157: 68/69)

(123) [Sales in the first half came to 159.92 billion yen, compared with 104.79 billion yen in the four-month period.]N

[Shiseido predicted that sales for the year ending next March 31 will be 318 billion yen, compared with 340.83 billion yen in the year ended Nov. 30, 1988…]S – Elaboration-additional (wsj_643: 11-12/13-20)

(124) [Most institutional investors have abandoned the portfolio insurance hedging technique, which is widely thought to have worsened the 1987 crash.]N [Not really insurance, this tactic was designed to soften the blow of declining stock prices and generate an offsetting profit by selling waves of S&P futures contracts…]S – Elaboration-general-specific (wsj_2381: 84-85/86-92)

2.2.3. (synonymy + subject NP)

Definition: A word (or phrase) in the satellite (also second) span is synonymous to (or is an acronym of) a certain word(s) present in the nucleus (also first) span, and that word (or phrase) in the satellite span is used as (i) the head of the subject NP of the sentence the satellite span starts with, or (ii) the head of the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(125) [As expected, Kellogg reported lower third-quarter earnings.]N [Net fell 16% to $123.1 million, or $1.02 a share, from $145.7 million, or $1.18 a share.]S – Elaboration-additional (wsj_610: 96-97/98)

(126) [Mr. Antar was charged last month in a civil suit filed in federal court in Newark by the Securities and Exchange Commission.]N [In that suit, the SEC accused Mr. Antar of engaging in a "massive

financial fraud" to overstate the earnings of Crazy Eddie, Edison, N.J., over a three-year period…]S – Elaboration-general-specific (wsj_File4: 77-78/79-89)

2.2.4. (meronymy + subject NP)

Definition: A set of objects or entities is introduced in the first (also nucleus) span, and a member object or entity from that set is mentioned in the satellite span, and used as (i) the head of the subject NP of the sentence the satellite span starts with, or (ii) the head of the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(127) [Some lagging competitors even may leave the personal computer business altogether.]S [Wyse Technology, for instance, is considered a candidate to sell its troubled operation…]S – Example (wsj_2365: 96/97-108)

(128) [Total sales gained 20% to 122.36 billion yen from 102.01 billion yen.]N [Exports made up 46.2% of the latest year's total, up from 39.8% a year ago…]S – Elaboration-set-member (wsj_657: 17/18-20)

2.2.5. (general word + subject NP)

Definition: A general word (e.g., *thing, matter* and *issue*), referring to an object, entity, fact or proposition in the first or nucleus span is used as (i) the head of the subject NP of the sentence the satellite span starts with, or (ii) the head of the subject NP of the sentence the embedded sub-nucleus span inside the satellite span starts with. E.g.,

(129) [Some of the associations have recommended Dr. Alan D. Lourie, 54, a former patent agent with a doctorate in organic chemistry who now is associate general counsel with SmithKline Beckman Corp. in Philadelphia. Dr. Lourie says the Justice Department interviewed him last July.]N

[Their effort has received a lukewarm response from the Justice Department...]S – Elaboration-additional (wsj_601: 33-36/37-75)

(130) [Eastman Kodak Co., seeking to position itself in the potentially huge high-definition television market, unveiled a converter that can transform conventional motion-picture film into high-definition video.]N

[The move also helps the Rochester, N.Y., photographic giant ensure that its motion-picture film business… isn't made obsolete by the upstart HDTV business…]S – Elaboration-additional (wsj_1386: 1-4/5-13)

2.3. (lexical + syntactic) features

2.3.1. (indicative word + present participial clause)

Definition: The second span which is a present participial clause is preceded by an indicative word (e.g., *by, in*). E.g.,

(131) [House leaders had hoped to follow the Senate's lead]N [by getting an agreement from House committee chairmen...]S – Means (wsj_1963: 5/6-9)

(132) [In announcing the plant delay,]S [Kellogg Chairman William E. LaMothe said, "Cereal volume growth in the U.S. has not met our expectations for 1989."] – Circumstance (wsj_610: 46/47-48)

2.4. (syntactic + semantic) features

2.4.1. (parallel syntactic construction + lexical chain)

Definition: The spans (clausal segments) or part of the spans (phrasal segments) are parallel to each other in syntactic construction. The syntactic parallelism is also strengthened by the occurrence of lexical items present in a lexical chain between the spans. E.g.,

(133) [Imports rose 11% to 18.443 trillion lire in September from a year earlier,]N [while exports rose 17% to 16.436 trillion lire.]N – Comparison (wsj_615: 11/12)

(134) ["Aerospace orders are very good," Mr. Cole says.]N ["And export business is still good.]N – List (wsj_628: 18-19/20)

2.5. (syntactic + positional) features

2.5.1. (present participial clause + beginning)

Definition: The satellite span which is a present participial clause is used in the beginning of the sentence containing both spans. E.g.,

(135) [Seeing all those millions in action,]S [I was just so relieved that Ms. Gruberova, gawky thing that she is, didn't accidentally smother herself in a drape.]N – Circumstance (wsj_1154: 42/43-46)

(136) [Commenting on the results for the quarter,]S [Mr. Treybig said the strength of the company's domestic business came as "a surprise" to him,]N – Circumstance (wsj_2396: 16/17-18)

2.5.2. (past participial clause + beginning)

Definition: The satellite span which is a past participial clause is used in the beginning of the sentence containing both spans. E.g.,

(137) [Led by its oat-based Cheerios line,]S [General Mills has gained an estimated 2% share so far this year, mostly at the expense of Kellogg.] – Circumstance (wsj_610: 16/17)

(138) [Cast as Violetta Valery in a new production of Verdi's "La Traviata,"]S [Ms. Gruberova last week did many things nicely and others not so well.]N – Elaboration-additional (wsj_1154: 7/8)

2.6. (graphical + syntactic) features

2.6.1. (comma + present participial clause)

Definition: The first span (usually the nucleus) is respectively followed by a comma and a present participial clause which is the second span (usually the satellite). E.g.,

(139) [Exxon Corp. filed suit against the state of Alaska,]N [charging state officials interfered with the oil company's initial efforts to treat last spring's giant oil spill.]S – Manner (wsj_1347: 1/2-4)

(140) [John J. Crabb sold 4,500 shares for $11.13 each,]N [leaving himself with a stake of 500 shares.]S – Elaboration-additional (wsj_1157: 82/83)

2.6.2. (comma + past participial clause)

Definition: The first span (usually the nucleus) is respectively followed by a comma and a past participial clause which is the second span (usually the satellite). E.g.,

(141) [Only a few months ago, the 124-year-old securities firm seemed to be on the verge of a meltdown,]N [racked by internal squabbles and defections.]S – Elaboration-additional (wsj_604: 2/3)

(142) [The dollar finished softer yesterday,]N [tilted lower by continued concern about the stock market.]S – Circumstance (wsj_1931: 1/2)

3. Unsure

Unsure refers to those cases in which no potential signals were found or were specified. E.g.,

(143) [This hasn't been Kellogg Co.'s year.]S [The oat-bran craze has cost the world's largest cereal maker market share.]N – Cause (wsj_610: 1/2)

(144) [All that now has changed.]N ["We're ahead for the year because of Friday," said the firm's Kurt Feshbach. "We're not making a killing, but we had a good day."]S – Explanation-argumentative (wsj_2381: 125/126-130)

(145) [Ed Shea and Barbara Orson never find a real reason for their love affair as the foolish, idealistic young Vass and the tirelessly humanitarian doctor Maria Lvovna.]N [Cynthia Strickland as the long-suffering Varvara is a tiresome whiner, not the inspirational counterrevolutionary Gorky intended.]N – List (wsj_1163: 95/96-97)

(146) ["This is a democratic process]N [-- you can't slam-dunk anything around here."]N – Consequence (wsj_1963: 33/34)

9. Appendix C

List of Discourse Markers

In our corpus analysis, we have identified 201 different DMs, as shown in Table 5.

#	Discourse Marker	#	Discourse Marker	#	Discourse Marker
1	Accordingly	68	Either/or	135	Most importantly
2	Additionally	69	Elsewhere	136	Most of all though
3	Admittedly	70	Essentially	137	Naturally
4	After	71	Even after	138	Nevertheless
5	After all	72	Even as	139	Nonetheless
6	After that	73	Even before	140	Nor
7	Afterwards	74	Even before then	141	Normally
8	All of a sudden	75	Even if	142	Not only/but
9	Along the way	76	Even now	143	Now
10	Already	77	Even so	144	Obviously
11	Also	78	Even though	145	Oddly
12	Alternatively	79	Even when	146	Of course
13	Although	80	Even while	147	On the contrary
14	And	81	Even with	148	On the other hand
15	And after	82	Even without	149	Once
16	And as a result	83	Eventually	150	Only if
17	And especially	84	Everytime	151	Only when
18	And even then	85	Evidently	152	Operationally
19	And for now	86	Except	153	Or
20	And for that reason	87	Except that	154	Otherwise
21	And now	88	Except when	155	Overall
22	And simultaneously	89	Finally	156	Particularly
23	And since then	90	For	157	Particularly as
24	And so	91	For example	158	Predictably
25	And still	92	For instance	159	Previously
26	And subsequently	93	For now	160	Quite the contrary
27	And then	94	For one	161	Rather
28	And thereby	95	For one thing	162	Rather than
29	And therefore	96	Fortunately	163	Recently
30	And thus	97	Further	164	Regardless
31	And unfortunately	98	Furthermore	165	Right now
32	And yet	99	Generally	166	Rightly or wrongly
33	Anyway	100	Given	167	Sadly
34	As	101	Given that	168	Separately
35	As a result	102	Hence	169	Since
36	As a result of	103	However	170	Since then
37	As far as	104	Ideally	171	So
38	As if	105	If	172	So far
39	As long as	106	If/then	173	So that
40	As soon as	107	Immediately	174	Still
41	As though	108	In addition	175	Supposedly
42	Aside from	109	In addition to	176	Then
43	At least	110	In any case	177	Theoretically
44	At that point	111	In any event	178	Thereafter

#	Discourse Marker	#	Discourse Marker	#	Discourse Marker
45	At the same time	112	In case	179	Therefore
46	At the time	113	In contrast	180	Though
47	Because	114	In essence	181	Thus
48	Because of	115	In fact	182	Typically
49	Before	116	In general	183	Ultimately
50	Besides	117	In other words	184	Unfortunately
51	But	118	In particular	185	Unless
52	But also	119	In response to	186	Until
53	But at the same time	120	In spite of	187	Until recently
54	But even	121	In the meantime	188	Until then
55	But even so	122	In this way	189	Upon
56	But eventually	123	In turn	190	Whatever
57	But in the end	124	Increasingly	191	When
58	But now	125	Indeed	192	When and if
59	But since then	126	Instead	193	When then
60	But so far	127	Instead of	194	Whenever
61	But then	128	Ironically	195	Where
62	By contrast	129	Irrespective of	196	Whereas
63	By the way	130	Just when	197	Whereby
64	Certainly	131	Meanwhile	198	While
65	Consequently	132	More importantly	199	With
66	Currently	133	More provocatively	200	Without
67	Despite	134	Moreover	201	Yet

Table 5: List of discourse markers