TECHNICAL SPECIFICATION

ISO/TS 24617-5

First edition
2014-03-01

Language resource management — Semantic annotation framework (SemAF) —

Part 5: Discourse structure (SemAF-DS)

Gestion de ressources langagières — Cadre d'annotation sémantique (SemAF) —

Partie 5: Structures de discours (SemAF-DS)

Reference number
ISO/TS 24617-5:2014(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

The committee responsible for this document is ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 04, *Language resource management*.

ISO 24617 consists of the following parts, under the general title *Language resource management — Semantic annotation framework*:

— *Part 1: Time and events (SemAF-Time, ISO-TimeML)*

— *Part 2: Dialogue acts (SemAF-DA)*

— *Part 4: Semantic roles (SemAF-SR)*

— *Part 5: Discourse structures (SemAF-DS)*

— *Part 6: Principles of semantic annotation (SemAF-Basics)*

— *Part 7: Spatial information (ISO-Space)*

— *Part 8: Semantic relations in discourse (SemAF-DRel)*

Introduction

Discourse structures play an essential role in the production and analysis of the syntactic, semantic, and pragmatic features of text, speech, and other types of discourse. This Technical Specification is a basis both for the annotation, generation and translation (among other processes) of these types of discourses and of the syntactic, semantic, and pragmatic features derived from them. Note that discourse structures underlie not only verbal communication (whether spoken, written, or signed) but also non-verbal discourse (such as a silent video).

The annotation scheme provided here specifies discourse structures that consist of segment structures and content structures. It also specifies the mappings between these two structures; the mappings are described by the annotations of discourse segments in texts or some other modalities. In this context, on the one hand, segment structures are spatiotemporal relations that hold between surface segments (such as words, phrases, clauses, sentences, and video scenes) and, on the other hand, content structures are discourse relations that are established between semantic and pragmatic items. Both of these structures can be represented by means of labelled directed graphs or sometimes simply by trees, as standardized by LAF (ISO 24612:2012) and SynAF (ISO 24615:2010).

This scheme also provides a common, language-neutral pivot for the interoperation among diverse formats of discourse structures of various types of document, and can be applied to the generation of linguistic and non-linguistic expressions. For example, if the discourse structures of speech and other linguistic data contained in motion pictures are fitted to this scheme, multilingual subtitles for these pictures can be generated at a reduced cost by means of a standardized tool for multilingual translation. By the same token, this scheme can facilitate interoperability among various discourse corpora and collaboration among researchers who use them.

Language resource management — Semantic annotation framework (SemAF) —

Part 5:
Discourse structure (SemAF-DS)

1 Scope

A discourse is a process of communication. This Technical Specification addresses how a discourse is structured in terms of its realization/presentation and content, and shows how its dual structure can be represented in a graph. The current specification focuses on the annotation of discourse structures in text only, but it can be extended to discourses in other modalities.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 15938-5:2003/Amd.1:2004, *Information technology. Multimedia content description interface. Part 5: Multimedia description schemes AMENDMENT 1: Multimedia description schemes extensions (MPEG-7 MDS AMD1)*

ISO 24612:2012, *Language resource management — Linguistic annotation framework (LAF)*

ISO 24615:2010, *Language resource management — Syntactic annotation framework (SynAF)*

ISO 24617-1:2012, *Language resource management — Semantic annotation framework — Part 1: Time and events (SemAF-Time, ISO-TimeML)*

ISO 24617-2:2012, Language resource management — *Semantic annotation framework — Part 2: Dialogue acts*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1
circumstance
entity which is an event (including dialogue act), state, process, relation, proposition, or set of these

3.2
class
unary predicate, which is a set of entities

3.3
discourse
process of communication, consisting of one or more sentences or sentence fragments

Note 1 to entry: From an abstract viewpoint, data (e.g. words, phrases, sentences, and paragraphs) representing a communication process is regarded as a discourse. A discourse can be encoded in various media such as text, hypertext, audio, video, and their possible combinations.

3.4
discourse relation
semantic/pragmatic relation that holds among two or more circumstances

Note 1 to entry: Some discourse relations, such as *example* and *part*, can also hold between objects. In this document, semantic/pragmatic relations (including discourse relations) are given in italics in the text and with a gray background in the Figures (e.g. *agent*, *inference*, and *purpose*).

3.5
discourse structure
structure of discourse, comprising segment structure, content structure, and possibly other types of structure

3.6
entity
semantic/pragmatic entity referenced in discourse, including circumstances, and objects

Note 1 to entry: An entity is represented by a node in a content structure.

3.7
object
semantic entity other than circumstance

Note 1 to entry: Objects include people, buildings, machines, ideas, and rules.

3.8
relational class
class whose instances are circumstances equivalent to relations

3.9
segment
word, phrase, clause, sentence, paragraph, section, chapter, or other partial realization of discourse

Note 1 to entry: A synonym is a 'discourse segment'. A segment references a semantic and/or pragmatic entity, which can be a semantic/pragmatic relation. Intrasentential segments are syntactic constituents such as words, phrases, and clauses. Segments might or might not be continuous: this is discussed in the definition of connectives.

4 Overview

A discourse structure consists of two types of structure: segment structure and content structure. A segment structure (extending intrasentential syntax) is a data structure that describes how a discourse has been organized from a formal syntactic perspective. It consists of

a) a set of segments (some partial realizations of discourse), and

b) the syntactic relations holding among them.

A content structure (extending intrasentential semantics) is a data structure that describes from a logical point of view how a discourse has been organized. It consists of

a) the set of semantic and pragmatic components referred to by the segments of a segment structure (that is, by some segments of some discourse), and

b) the logical relations established between these semantic representations. These two structures organize the whole structure of each discourse.

Both types of structure and content structures in particular, can be represented by means of a labelled directed graph. Various syntactic relations in a segment structure can, for instance, be captured by a tree (single-rooted graph). Discourse relations in a content structure can also be captured by a more general graph: The nodes in the graph stand for semantic and pragmatic components and the edges formalize the relations holding among them. In one way, a segment structure is to a discourse (or part of

it) what a syntactic structure is to a sentence (or a sub-sentential component), and a content structure is to a discourse (or part of it) what a semantic structure is to a sentence (or a sub-sentential component).

Rhetorical Structure Theory (RST)[4] assumes that discourse has a tree-like structure that can be regarded as an amalgamation of segment structures and content structures. Corpus annotation based on RST[2] considers segment structures involving markables, their annotations and, implicitly, some sort of content structures derived from them. Other corpus annotation initiatives such as the Prague Dependency Treebank[3] and the Penn Discourse TreeBank[6] follow essentially the same approach. By contrast Segmented Discourse Representation Theory (SDRT)[1] explicitly discusses content structures called Segmented Discourse Representation Structures (SDRSs), and with less commitment to segment structures and the mapping thereof.

By integrating these recent practices in fields such as formal linguistics, knowledge representation and corpus annotation, this Technical Specification provides an annotation scheme to partially specify the segment structures and the mapping from them to their corresponding content structures. For the sake of interoperability across different ISO standards such as LAF and SynAF, this annotation scheme has been made interoperable with practices concerning syntax and intrasentential semantics; this mapping from segment structures to content structures is therefore a straightforward extension of the mapping from syntactic structures to semantic structures, as addressed in many corpora, including the Penn TreeBank (PTB)[7] and PropBank[5].

As for sentences, parse trees describe their syntax, and logical forms represent their semantics. As for discourses, however, their syntax (i.e. their formal organization) and semantics (i.e. their content and logical organization) have been discussed in a more intertwined manner. For instance, most of the literature such as Reference [4] has regarded discourse relations as carrying both semantic and pragmatic information. This is inconvenient when one wants to focus on the semantic aspects of discourses, for instance, which can be the case when dealing with hypertexts, games and so on, which lack prefixed temporal order of presentation, and when discussing multiple (e.g. multilingual) presentations of the same semantic content.

To distinguish the realization/presentation and the content of a discourse and to address the mapping between them, this Technical Specification defines segment structures, content structures, and annotations to segments (discourse units) as part of segment structures. Segment structures represent the way in which the discourse is arranged, and consist of segments (e.g. words, phrases, clauses, sentences, paragraphs, sections, and chapters) together with the syntagmatic organization relations holding among them. Content structures represent the semantic and pragmatic content of discourses, and consist of nodes and links that represent entities referenced by segments. The main goal of this Technical Specification is to define an annotation scheme that concisely addresses segment structures, content structures and mappings between them. In other words, each segment annotated according to this scheme should represent a set of correspondences between segment structures and content structures.

A major basis of this Technical Specification is ISO/IEC 15938-5:2003/Amd.1:2004. This Technical Specification is mostly restricted to discourse structures, although the Linguistic DS also deals with predicate-argument structures and dialogue acts.

This Technical Specification addresses both the intrasentential and intersentential aspects of segment structures. The annotation of intrasentential aspects is compliant with ISO 24615:2010; that of both the aspects is consistent with the other two published parts ISO 24617-1:2012 and ISO 24617-2:2012. Their annotations and representations can be encoded according to ISO 24612:2012 as it supports labelled directed graphs.

5 Segment structure

A segment structure of a discourse addresses its syntactic organizations. This Technical Specification assumes that, not all, but some segment structures are represented as trees with their nodes representing discourse segments. If segment *S* (as a sequential data such as text and speech) has directed descendants (called 'daughters'), *S* is their concatenation. For instance, Figure 1 represents the segment structure of a discourse 'Tom left. It was late.' which consists of two daughters 'It was late.' and 'Tom left.'

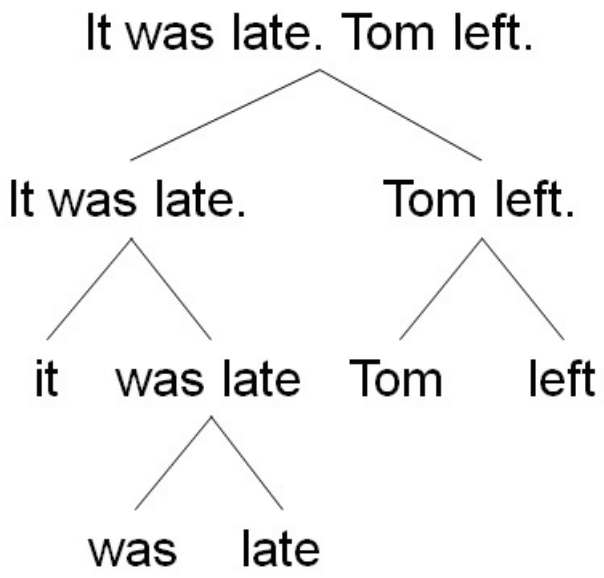

Figure 1 — Segment structure

A segment might, or might not, be continuous. For instance, 'either' plus 'or' in 'Either Tom is lying or Mary is mistaken' might be regarded as a discontinuous segment.

Daughters of a segment node in a segment structure may depend on one particular daughter of that node. Such a daughter is called a 'governing segment'; the others are called 'non-governing segments.' In this Technical Specification, a segment structure is encoded as a text containing annotations.

NOTE This Technical Specification is neutral between inline annotation and stand-off annotation because inline annotations are straightforwardly translated to stand-off annotations, as discussed in ISO 24612:2012.

By the conventions introduced here, a governing segment can be annotated by a pair of enclosing curly braces, and a non-governing segment by a pair of enclosing square brackets. This annotation may be partial in the sense that there can be segments without such markups.

In the following annotated sentence, for example, '{Tom left}' is a governing segment, '[{because} [it was late]]' a non-governing segment, '{because}' a governing segment, and '[it was late]' a non-governing segment. As such annotation is partial, neither 'Tom' is enclosed in square brackets, nor is 'left' enclosed in curly braces, for instance.

(1) [{Tom left} [{because} [it was late]].]

Below is an annotated discourse consisting of two sentences.

(2) [[It was late.] {Tom left.}]

Here, the first sentence (a non-governing segment) is regarded as dependent on the second sentence (a governing segment), so that the second is the nucleus of this discourse in the RST[4].

6 Content structure

Without loss of generality, semantic representations have been formulated as labelled directed graphs in formal semantics, knowledge representation (semantic network in particular) and related fields. Other types of semantic representation, such as logical forms and segmented discourse structures, can be translated to equivalent graphs. The current Technical Specification follows this practice and regard content structures of discourses as labelled directed graphs licensed by some ontology. All the nodes in a content structure are therefore typed by some classes in an ontology, and all the links there are

typed by some relations in that ontology. Like UML diagrams, the diagrammatic representation of those graphs elsewhere in this document is formal in the sense that they are not ambiguous or vague.

Content structures represent semantic (in a general sense, encompassing pragmatic) content of linguistic, non-linguistic, and other sorts of discourse segment. Each node therein represents a semantic entity, which is either a circumstance (e.g. event, state, relation, or proposition) or an object (a physical object such as a dog, or an abstract object such as love or a rule). Each link in a content structure represents a semantic relation (e.g. thematic role, discourse relation and communicative function) between the two entities represented by the two end points of the link. Since this Technical Specification concerns discourse structures, most links in the content structures in this document represent discourse relations, and the framed nodes accordingly represent entities that can be their arguments.

Below is an annotated segment followed by a corresponding content structure in Figure 2.

(3) [{Tom left} [{because} [it was late]].]

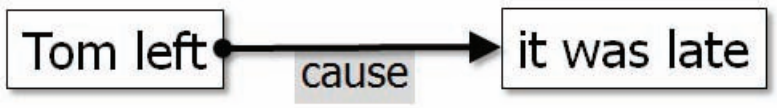

Figure 2 — Content structure corresponding to (3)

Each link in a content structure represents a semantic relation. The initial point and the terminal point of a link represent the first and the second argument of that relation, respectively. The *cause* link represents a *cause* relation between two arguments (note that this 'cause' is not a verb but a noun): The arrow points to the second argument, which is a cause of the first argument. In Figure 2, for instance, 'it was late' references a cause (the second argument of the *cause* relation) of the resulting event (the first argument of the *cause* relation) that 'Tom left' references.

Both of the framed nodes labelled 'Tom left' and 'it was late' in Figure 2 are abbreviated content structures of the respective segments. In Figure 3, the node labelled 'it was late' is an abbreviated content structure. However, balloon *E* in Figure 3 contains a non-abbreviated content structure of 'Tom left.'

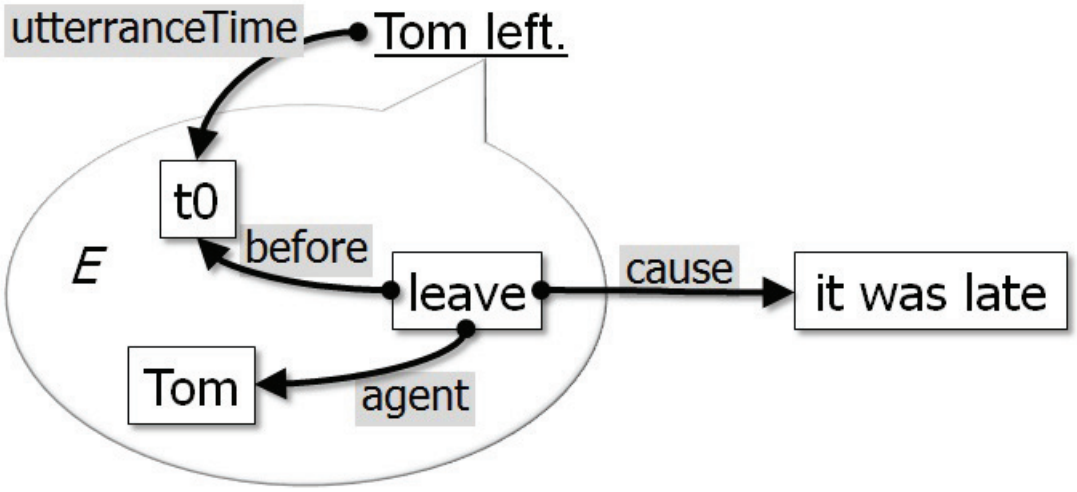

Figure 3 — Segment 'Tom left.' and detailed content structure

In general, if balloon *E* in Figure 3 is a content structure of segment *S* ('Tom left' in the current example), *E* may be abbreviated as a node *N* labelled by *S* (the framed node labelled by 'Tom left' in Figure 2). The unabbreviated content structure *E* is shown in Figure 3 by a balloon consisting of two links and three nodes. Here the *before* link conveys information that the leaving event took place before time t0, which is the utterance time of 'Tom left.' The *agent* link conveys Tom's being the agent of the leaving event. For

the sake of simplicity, the utterance time node (t0 in Figure 3) and the links connected with it will be disregarded throughout the rest of this document.

Note that the initial point of the *cause* link is the 'Tom left' node in Figure 2 but the `leave' node in Figure 3. This is because the 'leave' node constitutes the semantic core of the unabbreviated content structure *E* of segment, 'Tom left'. Such a node is usually called the head node of the segment. If a node *N* is an abbreviated content structure of segment *S* and is an end point of a link *L* like the *cause* link in Figures 2 and 3, the corresponding end point in an unabbreviated content structure *E* of *S* is its head node (the 'leave' node in Figure 3).

Some links are reifiable. A reifiable link can be reified to a node *N* together with two outgoing links to the two end points of the original link. The type of such a link must be a reifiable relation. In other words, a reifiable link of type *r* is an abbreviation of a node that is an instance of the relational class corresponding to *r* plus outgoing *arg0* and *arg1* links pointing to the first and the second argument of *r*, respectively. For example, the content structure in Figure 2 is an abbreviation of the following, where the *cause* relation in Figure 2 is a reifiable relation, the *cause* class in Figure 4 is the corresponding relational class, and the *cause* node represents an instance of that class.

Figure 4 — Content structure with reified link

Reifiable relations include discourse relations (e.g. *cause*, *causes*, *content* and, *conflict*) and semantic roles. However, relations such as *arg* (introduced below), *arg0* and *arg1* (see Figure 4) are not reifiable.

A reifiable relation may be modified or qualified. For instance, 'because' is modified by 'probably' in the sentence below:

(4) [{Tom left} [[probably] {because} [it was late]].]

This modification is captured by the content structure in Figure 5.

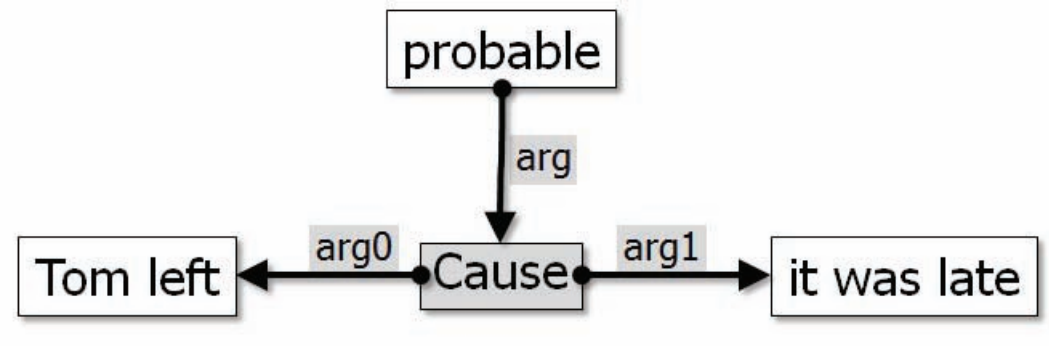

Figure 5 — Content structure of (4)

Here the *Cause* node represents the argument of the predication represented by the 'probable' node, meaning that the referenced *cause* relation is probable.

Content structures can contain hypernodes (graphs regarded as nodes). The segment structure in Figure 1 has the content structure in Figure 6, for example, which contains a hypernode enclosed in the gray frame. This hypernode consists of two nodes linked with each other and represents the content of the propositional attitude report.

(5) [Tom knows [that [[Bill loves Mary] but [she hates him]]].]

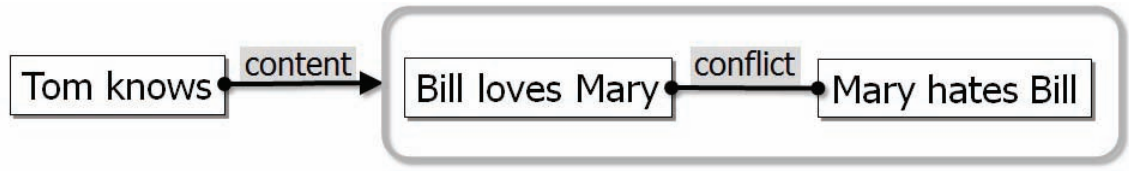

Figure 6 — Content structure corresponding to (5)

Hypernodes may represent scopes of propositional attitude reports. In Figure 6, the node 'Tom knows' is linked through a *content* relation, with the scope represented by the content structure consisting of two nodes 'Bill loves Mary' and 'Mary hates Bill' and a *conflict* link between them.

Conditionals, disjunctions, modal operators, quantifiers, negations and so on are also associated with scopes. For instance, the following conditional sentence has the content structure in Figure 7, containing two scopes represented by two gray frames.

(6) [[{If} [Tweety is a bird]], {he can fly}.]

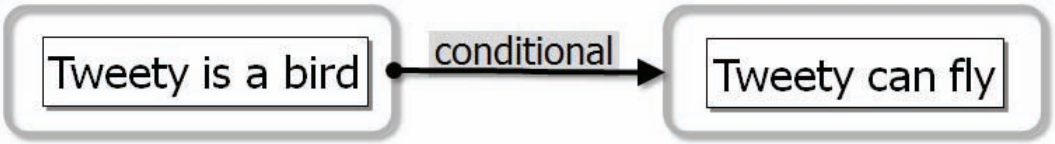

Figure 7 — Content structure corresponding to (6)

As discourses have such graph-based semantic representation, one would also expect to find graph-based semantic representation of clauses and sentences such as balloon *E* in Figure 3, but any other scheme of semantic representation may be used instead. In either case, ISO 24617-2:2012 and ISO/DIS 24617-4:2013 are recommended for a description of the internal structure of circumstances consisting of dialogue acts, communicative functions, predicate-argument structures and so on.

7 Mapping between segment and content structures

This clause discusses a mapping relation between segment structures and content structures. As illustrated in Figure 8, each segment structure S is associated with its content structure E. Here E contains two characteristic nodes: the head node and the governor node of S.

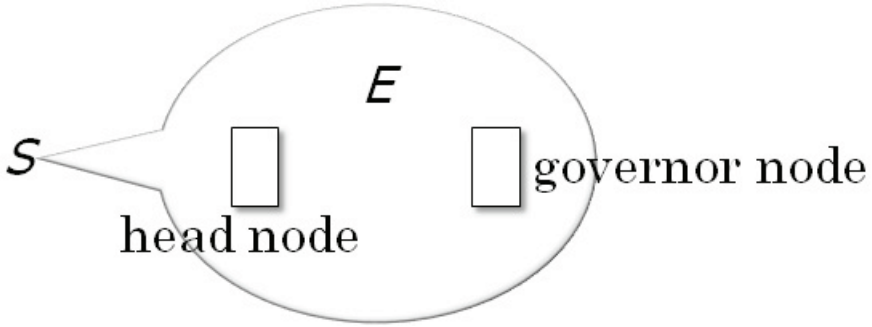

Figure 8 — Relation among segment structure S, content structure E, head node and governor node

As discussed in Clause 6, the head node of *S* is the node in *E* that forms the semantic core of *E*. The governor node of *S* is the head node of the segment that *S* depends on. *E* is the union of the content structures of S's daughters. *E* is therefore determined by the dependency relations among *S*'s daughters, given those daughters' content structures.

To summarize, the content structure *E* of segment *S* and those of *S*'s daughters are related as follows:

— *E* is the union of the content structures of *S*'s daughters;

— *S* and its head daughters share the same head node and the same governor node;

— the governor node of each daughter of *S* depending on *S*'s head is the head node of *S*.

Here, *S* is assumed to have at least one (possibly empty) head daughter. *S* can have multiple heads if it constitutes a coordinate construction, but this Technical Specification does not deal with such cases.

The above rules address the composition of content structures based on dependency. A segment *S* can optionally carry, in addition to the bracketing, the following three sorts of annotation: capturing the types of discourse relations, anaphora/coreference, and scoping.

a) A binary relation or a class, both encoded by texts with a gray background (such as *cause* and *causes*) in the beginning of *S.*

   1) When a binary relation *r* is used, *S*'s governor node *G* is the initial point of a link *l* of type *r*.

      i) If *r* is reifiable, *S*'s head node *H* is the instance of the relational class corresponding to *r* and *H* points to *G* via an *arg0* link (*H* and this *arg0* link are subsumed by link *l*).

      ii) If *r* is not reifiable, *H* is the terminal point of *l*.

   2) When a class *R* is used, *H* and *G* are the same instance of *R*.

b) The terminal point of *l* if a binary relation is used above, or the head node of *S* if a class is used above This is encoded by a subscript to *S* in this document.

c) The scope containing the referent of *S*: This is encoded by a superscript to *S* in this document.

According to item a) 1) i), the segment '{*cause* because}', for example, (in 'Tom left because it was late,' 'Tom left because of the accident.' and so on), has the following unabbreviated content structure, because the *cause* relation (*r*) is reifiable.

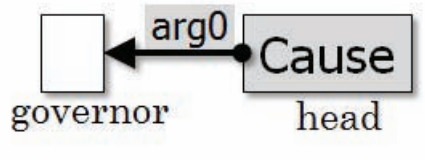

Figure 9 — Content structure of '{*cause* because}'

Here, the head node *H* in item a) is the *cause* node, which is the instance of relational class *cause*, corresponding to the *cause* relation *r*. This *H*, the *cause* node, points to the governor node (the empty frame) through the *arg0* link.

According to item a)1) ii), the segment '[*arg1* it was late]' (in 'Tom left because it was late' and so on) has the following content structure, since *arg1* is not reifiable.

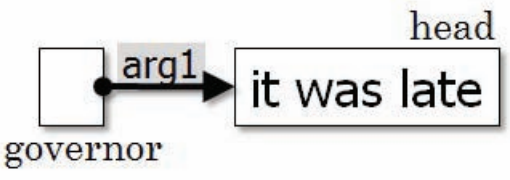

Figure 10 — Content structure of '[*arg1* it was late]'

Here, the head node *H* is the 'it was late' node. This *H* is the terminal point of the *arg0* link (*l*).

According to item a) 2), the segment '{*cause* cause}' (for instance in 'The cause of the accident was the tsunami.') has the following content structure containing just one node, which is the head node *H* and the governor node *G*.

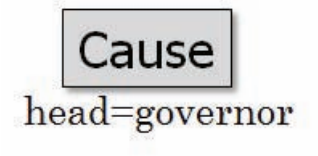

Figure 11 — Content structure of '{*Cause* cause}'

According to item b), the segment '[$_L$ *cause* So]' (in 'So Tom left' among others) has the content structure set out below: in it, *L* is the terminal node of *l* (the *cause* link reified in Figure 12).

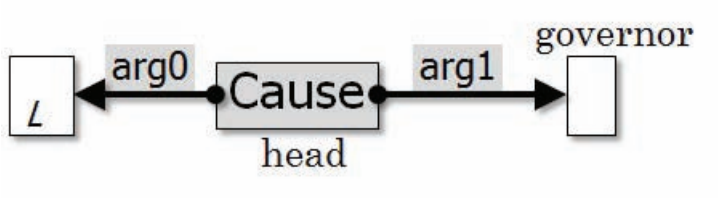

Figure 12 — Content structure of '[L cause S0]'

According to item c), '[^{X} Tom left]' entails that Tom's leaving event is in the scope *X*.

A connective is a segment referencing a semantic relation. This Technical Specification only discusses discourse connectives, that is to say connectives referencing discourse relations. There are two types of connective: dependent and independent. A dependent connective *C* (e.g. 'because') is a connective whose governor node represents the first argument of the semantic relation referenced by *C*, as in Figures 9 and 10. An independent connective *C* (e.g. 'cause') as a noun is a connective whose head node and governor node are both equal to an instance of the relational class corresponding to the relation referenced by *C*, as shown in Figure 11. Following item a) 2), the semantic-type annotation to an independent connective is hence a class (unary predicate) rather than a binary relation

As a compound example, segment (7) and its subsegments have the content structures as shown in Figure 13, where each balloon associated with a segment is its content structure.

(7) [{Tom left} [{*cause* because} [it was late]].]

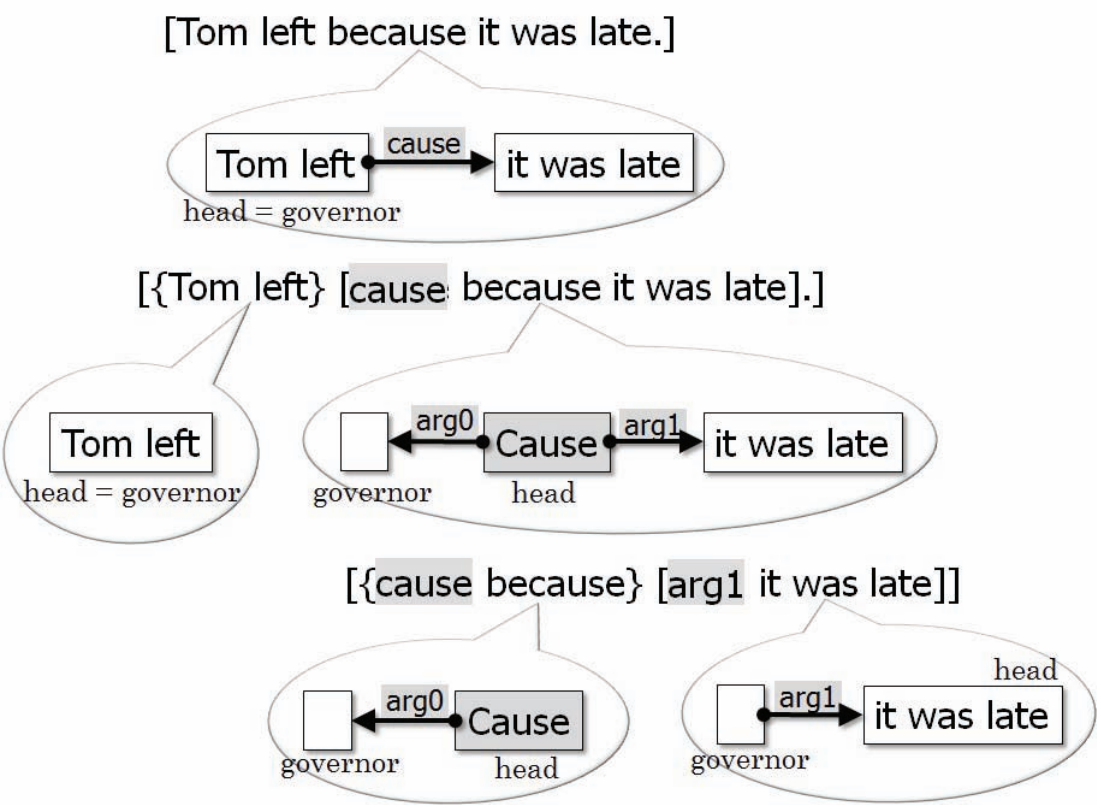

Figure 13 — Composition of content structure corresponding to (7)

The governor node of a complement segment *S* (e.g. 'it was late' as the complement of 'because') of a dependent connective C points to the head node of S through an *arg1* link. As 'because' governs 'it was late' in (7), the governor node of 'it was late' points to its head node through an *arg1* link, and the content structure of 'because it was late' therefore looks like the one in the right-hand balloon in the middle row of Figure 13. As 'Tom left' governs 'because it was late', the governor node of 'because it was late' equals to the head node of 'Tom left', and the content structure of 'Tom left because it was late' is therefore the one in the balloon at the very top of Figure 13.

Below is an example involving 'probably' modifying dependent connective 'because.'

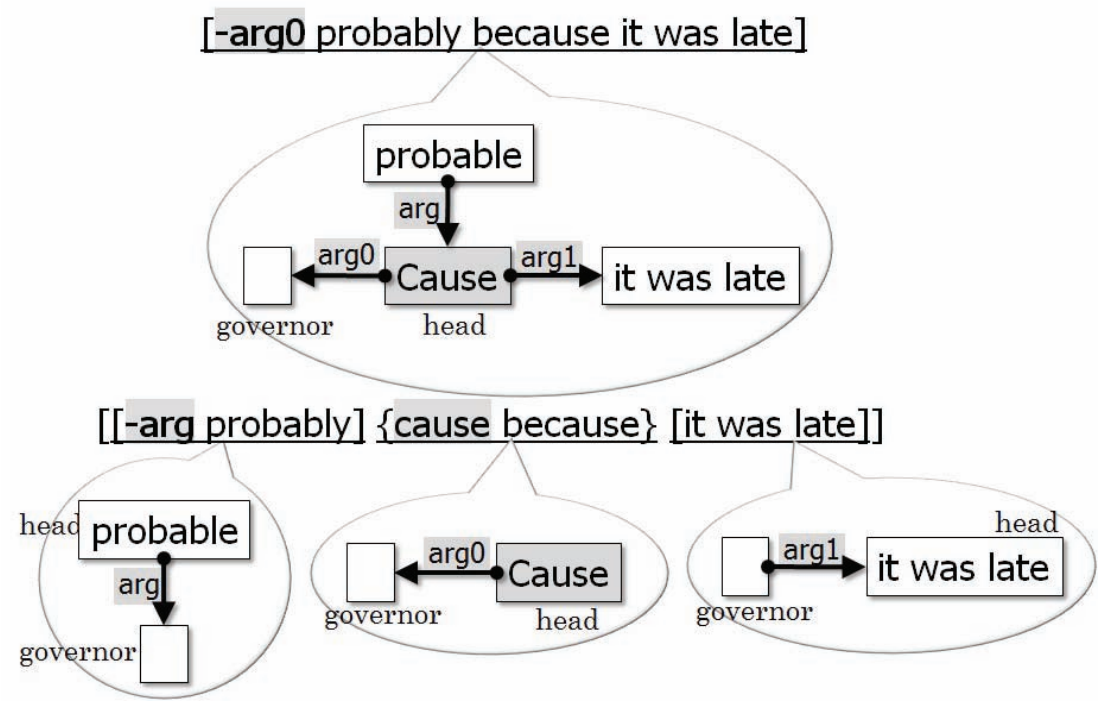

Figure 14 — Composition of content structure of 'probably because it was late'

Note that the governor node of 'probably' and 'it was late' is the *Cause* node, which is the head node of 'because'. If the governor node of 'because' and 'probably because it was late' is equal to the head node of 'Tom left', the resulting content structure is that in Figure 5. The minus sign in front of a relation name is the inverse operator; for example, *-arg* is the inverse of *arg*.

Below is an example of a multi-sentence discourse. The first sentence 'It was late' is regarded as containing an empty connective referencing a *cause* relation, and that its governor node therefore points to its head node through a *cause* link, according to item a) 1). As 'It was late' depends on 'Tom left', the governor node of the former is the head node of the latter, thereby generating the upper content structure of the whole discourse.

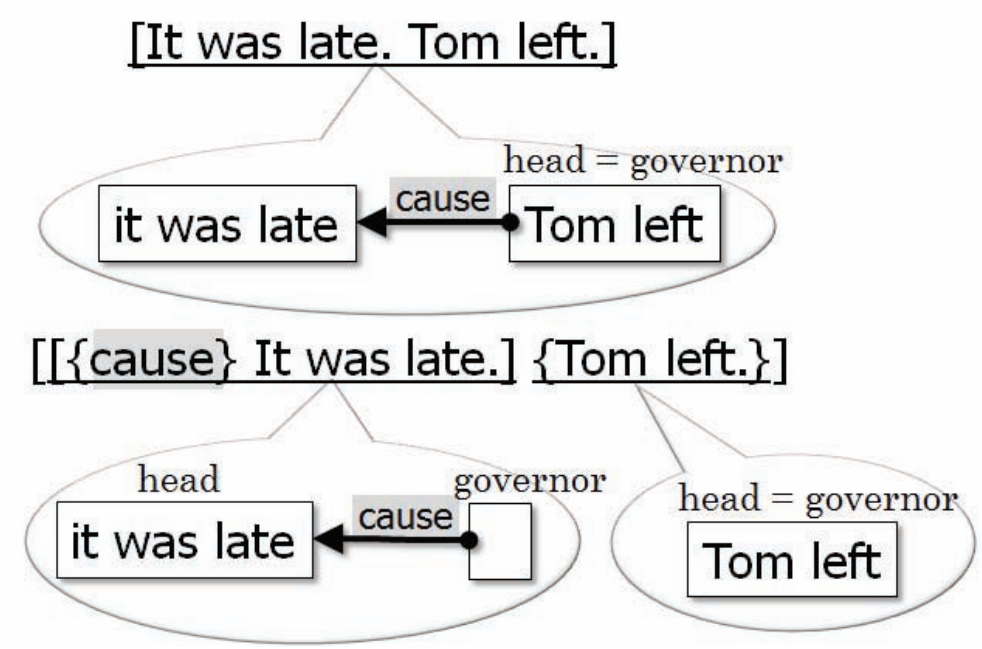

Figure 15 — Composition of content structure of 'It was late. Tom left.'

Below is another multi-sentence example involving an anaphoric dependent connective 'so'.

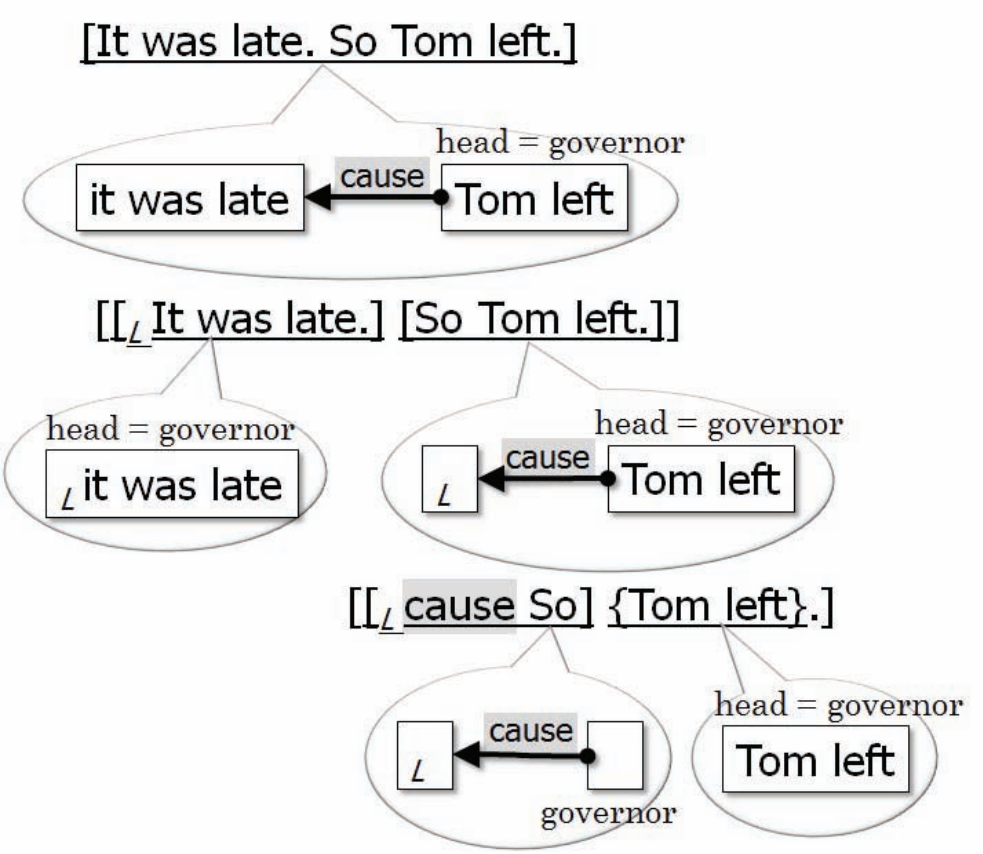

Figure 16 — Composition of content structure of 'It was late. So Tom left.'

According to item b), index *L* shared by '[$_L$ It was late.]' and '[$_L$ *cause* So]' means that 'it was late' is the antecedent of the zero anaphor that is contained in 'so', and references the second argument of the *cause* relation.

The head node of 'so' is a *Cause* node, which can be the governor node of 'probably', for instance, if 'probably' modifies 'so' as below:

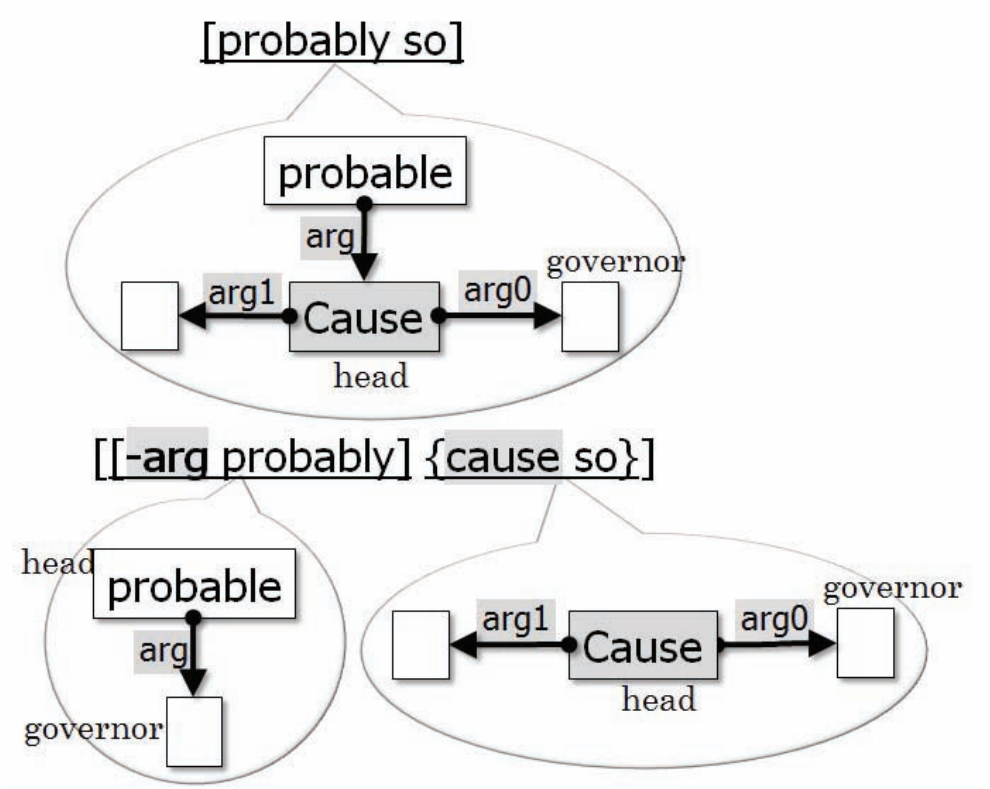

Figure 17 — Composition of content structure of 'probably so'

Below is the composition of a sentence containing an independent connective 'cause' as the main verb referencing a causing event, which in turn is represented by a *Causes* node together with an *arg0* link and *arg1* link as shown below, according to a).

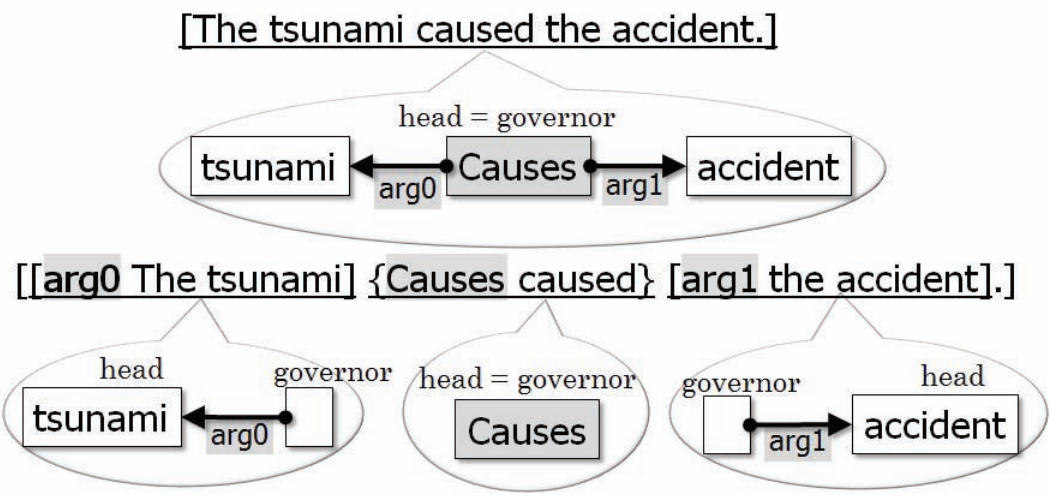

Figure 18 — Composition of content structure of 'The tsunami caused the accident.'

Here, the semantic representation of the past tense (as in Figure 3) is omitted for the sake of simplicity. The *causes* relation is the inverse of the *cause* relation, and the first argument of *causes* is therefore a cause of the second argument (which is the result).

The example below contains an independent connective 'cause' as a noun that references a *Cause* event, which is also the same as a *Causes* event except that the first argument and the second argument of *Cause* reference the result and the cause, respectively.

(8) [[[The] {*Cause* cause} [{*arg0* of} [the accident]] {is} [*arg1* the tsunami].]

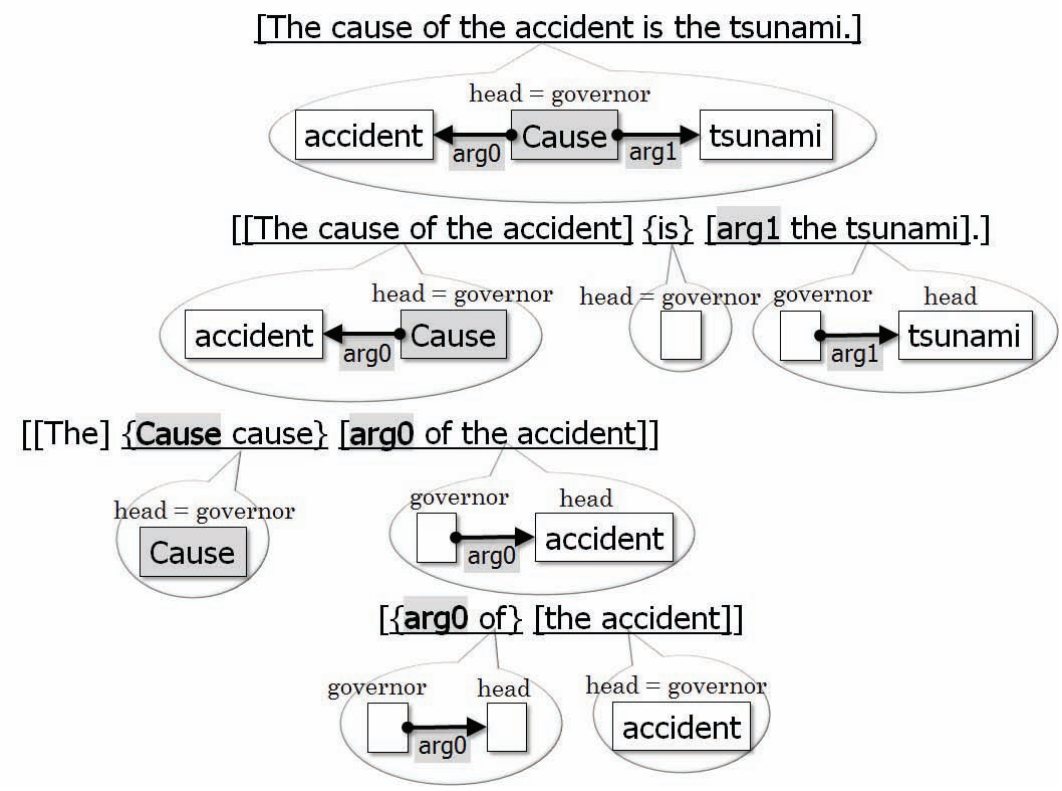

Figure 19 — Composition of content structure corresponding to (8)

The *Cause* node as the head node of 'cause' equals to the governor node of 'of this accident' to account for the content structure of 'The cause of the accident.' The head node (which is just a placeholder) of 'is' equals to the governor node (the *Cause* node) of 'The cause of this accident' and that of 'the tsunami' to generate the content structure of the entire sentence at the top of Figure 19.

In this connection, Figure 20 below shows how 'the apparent cause' is composed, which may replace the composition of segment structure and content structure of 'cause' in (8) and Figure 19, respectively.

(9) {[-*arg* apparent] {*Cause* cause}}

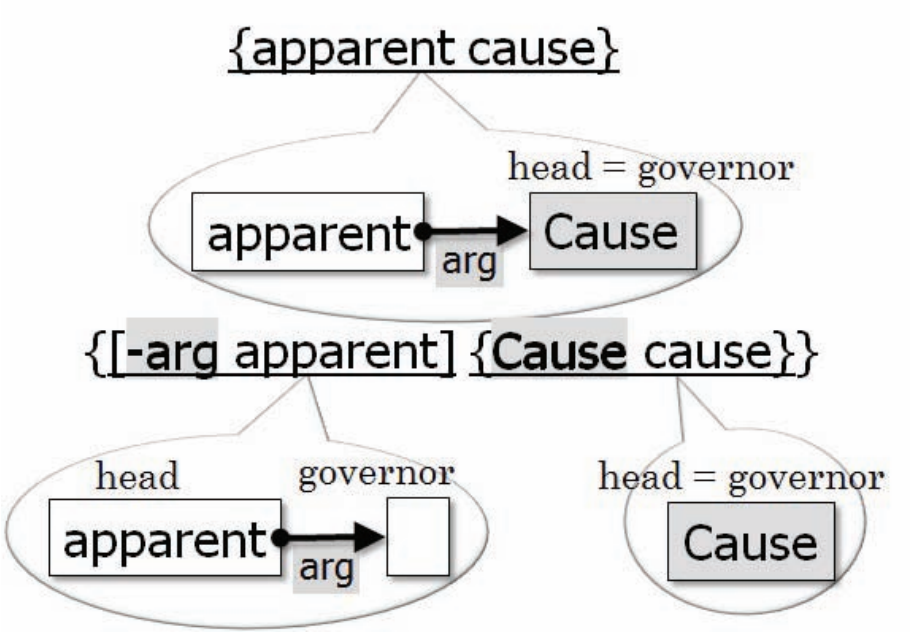

Figure 20 — Composition of content structure corresponding to (9)

Coordinate conjunctions such as 'and' and 'but' are independent connectives. For instance, the following composition involves a coordination and a sentence adverb.

(10) [[-*arg* Unfortunately], [[*arg0* Bill loves Mary] {*Conflict* but} [*arg0* she doesn't like him]].]

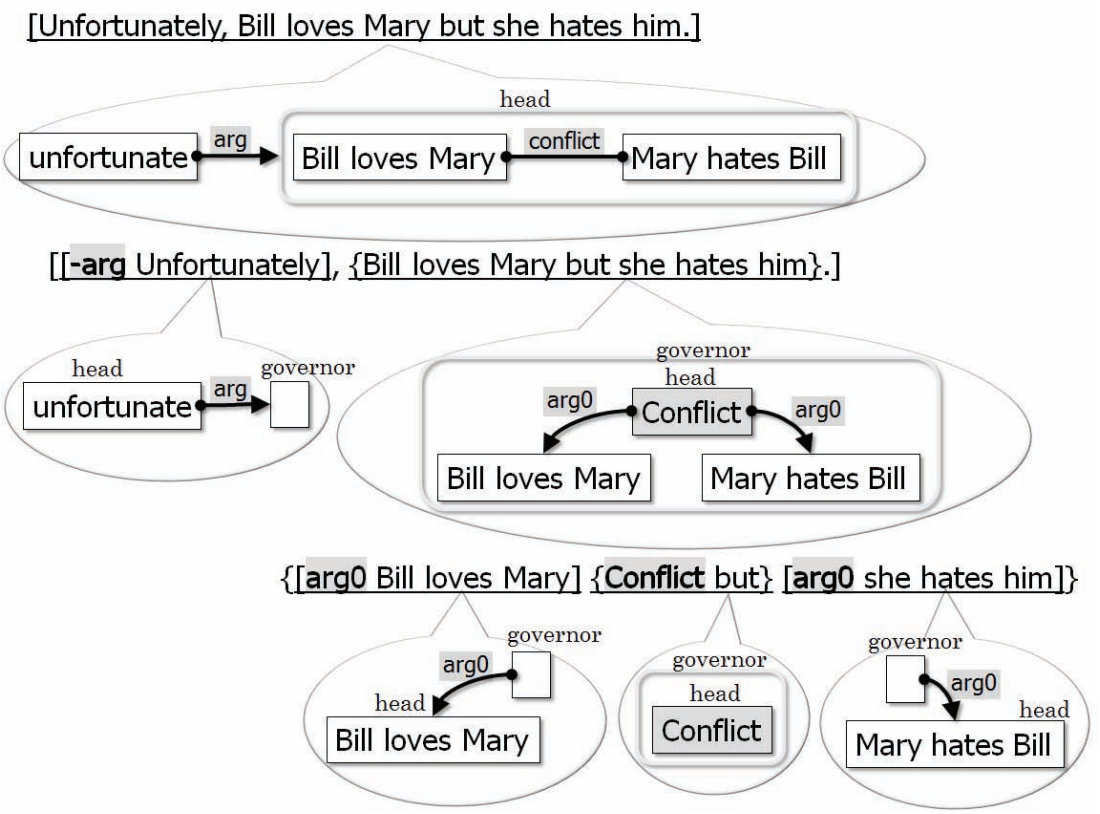

Figure 21 — Composition of content structure corresponding to (10)

If 'but' can only govern its complements, 'Bill loves Mary but she hates him' is a non-governing segment (maximal projection), and its governor node is therefore equal to the head node of the whole sentence, which is in turn equal to the governor node of 'unfortunately'.

Propositional attitude reports typically involve scopes; the following example is (5) with some annotations, where *content* is regarded as a non-reifiable relation:

(11) [Tom knows [{*content* that} [[Bill loves Mary] {*Conflict* but} [she hates him]]].]

The corresponding content structure is in Figure 6.

In (12), the referents of 'it's raining' and 'I don't have an umbrella' belong to scope *X*, which is the second argument of the *reason* relation referenced by 'So,' according to item c).

(12) [[$_X$ [^{X} It's raining.] [^{X} I don't have an umbrella.]] [[$_X$ *reason* So] {I'll stay here}.]]

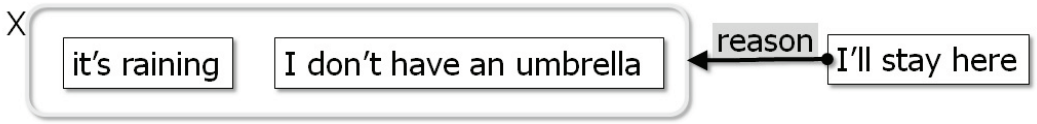

Figure 22 — Content structure corresponding to (12)

8 Concluding remarks

This Technical Specification has specified how to annotate discourse structures in terms of segment and content structures with mapping between them. It has also specified how to represent the two types of structure in graphs.

Bibliography

[1] Asher N., & Lascarides A. *Logics of Conversation*. Cambridge University Press, 2003

[2] Carlson L., Marcu D., Okurowski M.E. 2003) Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*, 85-112, Jan van Kuppevelt and Ronnie Smith (*eds*.), Kluwer Academic Publishers

[3] Hajič J. et al. *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia, 2006

[4] Mann W., & Thompson S. Rhetorical Structure Theory. A Theory of Text Organisation. *Text*. 1988, **8** (3) pp. 243–281

[5] Palmer M., Gildea D., Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput. Linguist.* 2005, **31** (1) pp. 71–105

[6] Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A. et al. 2008) The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*

[7] PTB. *The Penn Treebank Project*. http://www.cis.upenn.edu/~treebank/

ICS 01.020

Price based on 17 pages