First steps towards an ISO standard for annotating discourse relations

Harry Bunt*, Rashmi Prasad and Aravind Joshi*****

*Tilburg Center for Cognition and Communication, Tilburg University, The Netherlands
**Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, USA
***Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA
harry.bunt@uvt.nl, prasadr@uwm.edu, joshi@seas.upenn.edu

Abstract

This paper describes initial studies in the context of a new effort within ISO to design an international standard for the annotation of discourse with semantic relations that are important for its coherence, "discourse relations". This effort takes the Penn Discourse Treebank (PDTB) as its starting point, and applies a methodology for defining semantic annotation languages which distinguishes an abstract syntax, defining annotation structures as set-theoretical constructs, a concrete syntax, that defines a reference XML-based format for representing annotation structures, and a formal semantics. A first attempt is described to formulate an abstract syntax and a concrete syntax for the annotation scheme underlying the PDTB. The abstract syntax clearly shows an overall structure for a general-purpose standard for annotating discourse relations, while the resulting concrete syntax is much more readable and semantically transparent than the original format. Moreover, some additional elements are introduced which have an optional status, making the proposed representation format compatible not only with the PDTB but also with other approaches.

1. Introduction

With the recent availability of various types of linguistically annotated corpora developed for natural language processing (NLP), there is now an urgent need for addressing the demands for their *representational compatibility*, in order to ensure that each of these resources can be effectively merged, compared and manipulated with common software. An excellent example of the need for compatibility can be seen in the several different layers of annotations done on the Wall Street Journal (WSJ) corpus, such as POS tagging, syntactic constituency, coreference, semantic role labeling, events, and discourse relations. Although these annotations at different layers have resulted in a highly linguistically enriched corpus, efficient use of the resource for empirical NLP has been hindered by challenges in merging the linguistic data from the different levels because of their incompatible representations.

In addition to annotation representation, it is also necessary to ensure that when the same linguistic phenomenon is being annotated across different projects, each targeting a different language, domain, genre, or source text within the same genre, that this collective subcommunity agree on an *annotation schema standard* for the phenomenon. While agreement on schema standards is highly challenging to achieve, since it must be general enough to account for the full breadth of variation found across languages, domains, and genres, it is nevertheless necessary if we want to effectively utilize the collective resources for each phenomenon and move the state-of-the-art forward with big strides.

This work forms part of ISO efforts to establish international standards for semantic annotation. Two parts of the standard have so far been completed: ISO 24617-1 (Semantic annotation framework, Part 1: Time and events) and ISO 24617-2 (Semantic annotation framework, Part 2: Dialogue acts). Part 8, concerned with *relations in discourse*, was launched in 2011 and results from an agreement between the PDTB Research Group (`http://www.seas.upenn.edu/~pdtb`) and the ISO Working Group, ISO/TC 37/SC 4/WG 2 "Language resource management, Annotation and representation schemes", that a joint activity should take place to design an international standard for the annotation of discourse relations, taking the PDTB annotation scheme and guidelines (PDTB Group, 2008; Prasad et al, 2008) as the starting point. This work should include:

1. Adaptation of the PDTB annotation scheme as needed to conform to the requirements of ISO international standards;
2. Verification of the annotation scheme across a wide variety of languages, domains, and genres.

This paper describes preliminary studies for the first of these steps, in continuation of the work in Ide et al (2011). This part of ISO 24617 will provide definitions and representations of concepts for annotating explicit and implicit discourse relations. A notable feature of the abstract representation for the scheme is that it is designed to be flexible, to accommodate a certain degree of variation between schemes. This is implemented by means of optionality in the representation. Some novel concepts and structures are also introduced that are not represented in the current version of the PDTB.

2. The PDTB: A theory-neutral and lexically-grounded approach

The primary reason for adopting the PDTB as the basis for a discourse relation standard is that the framework avoids biasing the annotation towards any particular theory, and instead specifies discourse relations at a "low level" that is clearly defined and well understood. In particular, each relation, along with its two arguments, is annotated independently of other relations, and no further dependencies are shown among the relations. Thus, the argument structures annotated are strictly local. Since there is currently little agreement on a general theory of high-level discourse structure representation, with the proposed structures being variously trees, graphs, or DAGs (e.g., Hobbs, 1985;

Polanyi, 1987; Mann and Thompson, 1988; Webber et al., 2003; Asher and Lascarides, 2003; Wolf and Gibson, 2005; Lee et al., 2008) the theory-neutral approach of the PDTB should hold appeal for researchers across these theories, allowing for validation studies of the theories. In this sense, the PDTB framework provides a basis for an emergent and data-driven theory of discourse structure.

Another major appeal of the PDTB is its lexically-grounded approach to the annotation, lending to greater reliability of annotation, especially since it inferences at the level of discourse are much harder than at the sentence level.

The second (current) version of the PDTB, PDTB-2.0, is distributed through the Linguistic Data Consortium (LDC).[1]

3. Scope and Basic Concepts of PDTB

Discourse relations, such as causal, contrastive, and temporal relations, are considered to be semantic relations between abstract objects (eventualities and propositions), which are the arguments of the relation. The PDTB provides annotations of discourse relations, along with their arguments, senses and attributions, on the entire PTB-II portion of the WSJ corpus (Marcus, 1993), consisting of approximately 1 million words. In the rest of this section, we detail the basic concepts and elements of the PDTB annotation framework that underlie the proposed standard in this paper. It should be noted that the standards proposed here do not say anything about the overall annotation task design, workflows, and evaluation methods, for which the reader is referred to the PDTB reports and publications related to the annotation (Miltsakaki et al, 2004; Prasad et al., 2007; Miltsakaki et al., 2008; Prasad et al, 2008; PDTB-Group, 2008).

3.1. Discourse relations and their arguments

Discourse relations are often triggered by explicit words or phrases, such as the underlined expressions in (1a) and (1c), but they can also be implicit, as in (1b). Explicit realizations can occur via grammatically defined *(explicit) connectives* (1a), such as (subordinating and coordinating) conjunctions, adverbs and prepositional phrases, or with other expressions not so grammatically well-defined, called *Alternative lexicalizations* (*AltLex*) (1c). Each discourse relation is assumed to hold between two and only two abstract object (AO) arguments. Since there are no generally accepted abstract semantic categories for characterizing the arguments of discourse relations, they are simply labeled Arg1 (shown in italics) and Arg2 (shown in bold). For explicit connectives, Arg2 is the argument to which the connective is syntactically bound; Arg1 is the other argument.

(1) a. *Big buyers like P&G say there are other spots on the globe, and in India, where the seed could be grown* (...) But **no one as made a serious effort to transplant the crop.**

b. *Some have raised their cash positions to record levels.* Implicit=because **High cash positions help buffer a fund when the market falls.**

c. *But a strong level of investor withdrawal is much more unlikely this time around,* `fund managers said.` **A major reason is that investors already have sharply scaled back their purchases of stock funds since Black Monday.**

d. *Pierre Vinken, (...) will join the board as a nonexecutive director Nov. 29.* EntRel **Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.**

e. *Jacobs is an international engineering and construction concern.* NoRel **Total capital investment at the site could be as much as $400 million**

Between two adjacent sentences not related by an explicit connective or AltLex, an implicit discourse relation can be inferred, in which case the annotator has to *insert* a connective to express the inferred relation, such as the implicit connective *because* inserted in (1b). For such *(implicit) connectives*, the labels Arg1 and Arg2 reflect the linear order of the arguments (Arg1 occurs before Arg2).

Arguments of explicit connectives can be located anywhere in the text, whereas arguments of implicit connectives and AltLex must be adjacent. For either of these, there are no syntactic constraints to how far an argument can extend. Thus, arguments can be single clauses, sentences, or multiple clauses or sentences. From a semantic point of view, however, an argument must contain the *minimal* amount of text that is required for interpreting the relation. To facilitate the minimality-driven argument annotation, arguments are allowed to have *supplementary* text associated with them. A supplementary text annotated for an argument — Sup1 for Arg1 and Sup2 for Arg2 — indicates that this text was perceived as relevant (but not necessary) to the interpretation of the argument. Example 2(a) shows a Sup2 annotation (enclosed in square brackets) from the PDTB, where the explanation provided for the "suing" is considered to be relevant to Arg2 but not *necessary* to interpret the temporal relation expressed with "then".

(2) a. *It acquired Thomas Edison's microphone patent* **and** then **immediately sued the Bell Co.** [claiming that the microphone invented by my grandfather, Emile Berliner, which had been sold to Bell for a princely $50,000, infringed upon Western Union's Edison patent.]

It is also possible for adjacent sentences in a coherent discourse to not be related by any discourse relation, in particular when the sentences are linked by an entity-based coherence relation (*EntRel*, as in (1d)), or are not related at all via adjacency (annotated as *NoRel*, shown in (1e)). Arguments of EntRel relations must be adjacent to each other and cannot contain sub-sentential spans, although they can be extended to include multiple sentences. Arguments of NoRel are like EntRel except that the adjacent sentences cannot be extended to include additional sentences.

[1] `http://www.ldc.upenn.edu`, Entry LDC2008T05.

3.2. Senses of discourse relations

In the PDTB, senses of discourse connectives are represented in a flexible manner, via a three-tiered hierarchical classification going from four coarse-grained senses at the top *class* level to more refined meanings at the second *type* and third *subtype* levels. The full PDTB sense hierarchy is shown in Fig. 1. In the process of annotation, annotators can back off to the more coarse-grained levels when they have low confidence on the more refined senses. This is beneficial for achieving inter-annotator reliability, especially if agreement among annotators is measured in terms of a *weighted kappa* statistic (Geertzen and Bunt, 2006), which takes into account that a tag T_1 at one level and a tag T_2 at a lower level, such that T_2 is dominated by T_1, correspond to interpretations which are not identical and hence not fully in agreement, but which are in partial agreement. Annotations could also be carried out with just the *class* level or the *class* and *type* levels while ignoring the lower level senses.

The examples in (3) illustrate the use of sense tags in the PDTB to define a specific discourse relation. Sense tags are shown in parentheses, with the colon used to illustrate the hierarchical organized sense label when the most refined subtype sense was chosen (CLASS:TYPE:SUBTYPE).

(3) a. *Big buyers like P&G say there are other spots on the globe, and in India, where the seed could be grown ...* But **no one as made a serious effort to transplant the crop.** (Comparison:Concession:Contra-expectation)

b. *Some have raised their cash positions to record levels.* Implicit=because **High cash positions help buffer a fund when the market falls.** (Contingency:Cause:Reason)

c. *But a strong level of investor withdrawal is much more unlikely this time around,* `fund managers said.` **A major reason is that investors already have sharply scaled back their purchases of stock funds since Black Monday.** (Contingency:Cause:Reason)

Discourse connectives can be ambiguous, for example *since* has a temporal sense in (4a) but a causal sense in (4b). In such cases, annotation simply involves choosing the intended sense. But connectives can also have multiple senses. For example, *since* in (4c) has both the temporal as well as the causal sense. To handle multiplicity, multiple sense tags per connective must be allowed. In the PDTB, up to two senses per connective are admitted.

(4) a. *The Mountain View, Calif., company has been receiving 1,000 calls a day about the product* since **it was demonstrated at a computer publishing conference several weeks ago.**

b. *It was a far safer deal for lenders* since **NWA had a healthier cash flow and more collateral on hand.**

c. *Domestic car sales have plunged 19%* since **the Big Three ended many of their programs Sept. 30.**

Multiplicity needs to be allowed for implicit relations as well. This is implemented by allowing multiple implicit connectives to be inserted for an implicit relation, with each connective expressing one of the two inferred senses.

The PDTB sense hierarchy contains 43 sense tags, which form the total set of discourse relations distinguished in the PDTB. This reflects the idea that there is a rather small core set of semantic relations that can hold between the situations described in the arguments of connectives (Kehler, 2002). However, the core set of relations corresponding to the 'class' level can be refined by adding other types and subtypes, and can be viewed as an open set of possible relations. The use of a hierarchically organized set of 43 discourse relations makes a basic difference between the PDTB and RST-style labeling of discourse relations (Mann and Thompson, 1988).

3.3. Attribution

In the PDTB, each discourse relation, whether expressed explicitly by a connective, explicitly by alternative means, or implicitly by adjacency, and each of its arguments is annotated for *attribution*, i.e. for the source to whom the relation or an argument are ascribed, such as the author(s) (or speaker) of the text, as in example (5a), or someone else who is quoted in the text, as in example (5b). Preliminary studies for the PDTB have indicated that a substantial proportion (34%) of the annotated discourse relations have another source than the author of the text, either for the relation or for one or both of its arguments.

(5) a. Since **the British auto maker became a takeover target last month,** *its ADRs have jumped about 70%.*

b. *"The public is buying the market* when **in reality there is plenty of grain to be shipped"**, said Bill Biedermann, Allendale Inc. director.

The PDTB annotation scheme distinguishes four properties of attributions, which are annotated as feature specifications: *source, type, scopal polarity*, and *determinacy*.

The *source* of an attribution distinguishes between (a) the writer of the text ("Wr"); (b) some specific other agent introduced in the text ("Ot"); and (c) some arbitrary agent indicated in the text through a non-specific reference ("Arb").

The *type* of an attribution encodes the nature of the relation beween the agent who is the source of a discourse relation and the arguments of the relation. The following kinds of relation are distinguished: (a) communication (annotated as "Comm") for asserted relations, typically involving verbs like *say, claim, argue, explain*; (b) propositional attitude ("PAtt") for cases where the source expresses a belief, expectation, assumption, etc.; (c) factive ("Ftv") for cases where the source has indicated a relation to a certain fact, e.g. by using a verb like *regret, forget, remember,* or *see*; and (d) control ("Ctrl"), for a relation to an eventuality as expressed by a control verb like *persuade, permit, promise, want*, etc.

The *scopal polarity* of an attribution serves to identify cases where verbs of attribution are negated on the surface, but where the negation in fact reverses the polarity of the attributed relation or argument, as in example (6):

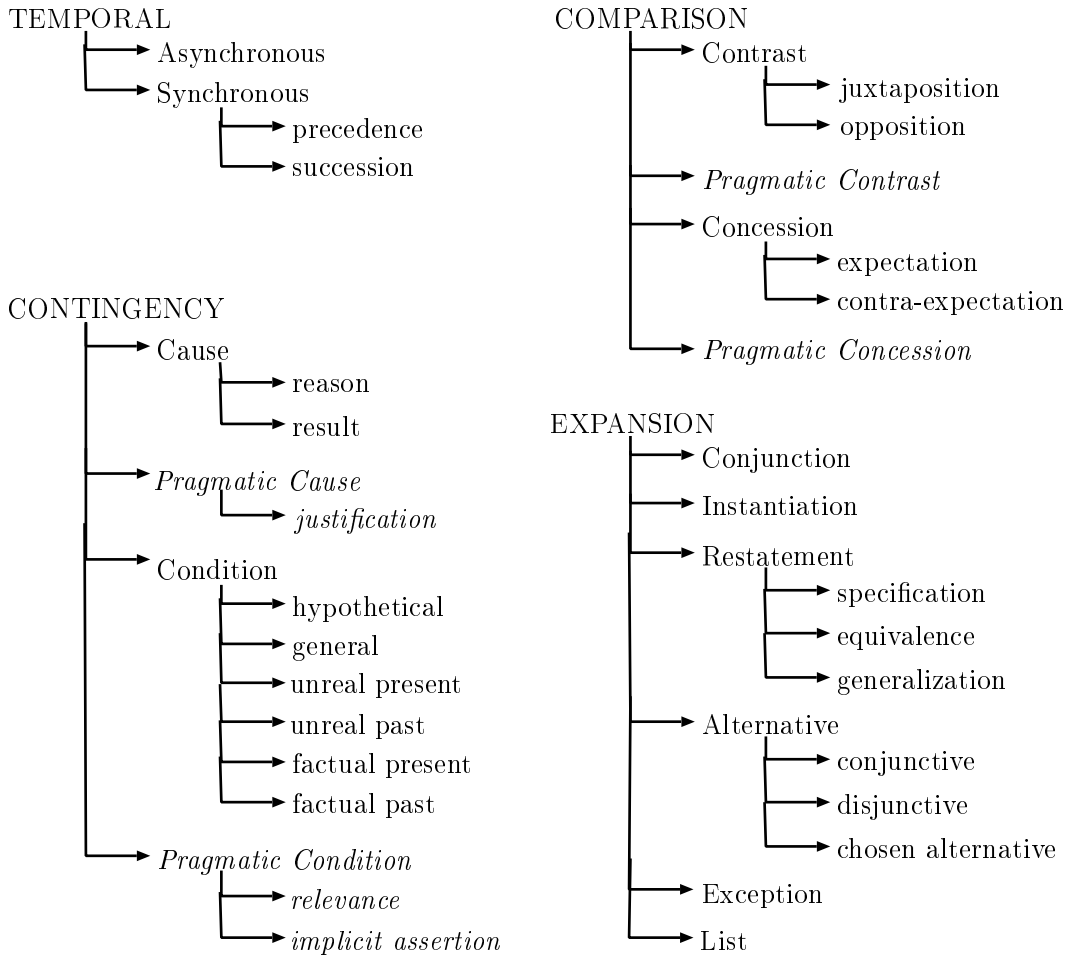

Figure 1: Hierarchy of discourse relations in the PDTB ('sense tags')

(6) *"Having the dividend increases is a supportive element in the market outlook,* but I don't think **it's a main consideration"**, he says.

In such cases, the relation (or the argument, as the case may be) is marked as having scopal polarity "Neg". This may occur both for explicit discourse relations expressed by a connective and for implicit relations.

The *determinacy* of an attribution is used to capture that the attribution may be cancelled or made indeterminate ("Ind") within a particular context, such as within the scope of a conditional or an infinitival, as in example (7), where the idea that "our teachers would educate our children better if only they got a few thousand dollars more" is not a belief or an opinion that is attributed to anyone.

(7) Its is silly libel on our teachers to think *they would educate our children better* if only **they got a few thousand dollars a year more.**

If there is no indeterminacy associated with an attribution, its determinacy has the default value "Null".

3.4. Representation format

In line with ISO requirements, PDTB annotations are in stand-off format: files containing the annotations are physically separate from the source text files. The PDTB annotation scheme and representation are fully described in the manual (PDTB-Group, 2008).

The current scheme for annotating a discourse relation entity in the PDTB includes a list of values, which may also represent text spans, as references to the character offsets in the source text file, and the PTB alignments of the text spans, as gorn address references to nodes in their corresponding PTB constituency trees. This may have to be revised in order to be ISO-compliant, following the joint ISO-TEI standard 24610-1 (see ISO 24610-1, 2006).

4. DRelML: Discourse Relations Markup Language

4.1. Overview

The Discourse Relations Markup Language DRelML has been designed in accordance with the ISO Linguistic Annotation Framework (LAF, ISO 24612:2009), which draws a distinction between the concepts of *annotation* and *representation*. The term 'annotation' refers to the linguistic information that is added to regions of primary data, independent of the format in which the information is represented; 'representation' refers to the format in which an annotation is rendered, independent of its content. According to LAF, *annotations* are the proper level of standardization, rather than representations. Conforming to the annotation-representation distinction, the DRelML specification fol-

lows the methodology for designing annotation languages developed in Bunt (2010), which has become standard practice in ISO work on semantic annotation. According to this methodology, the definition of an annotation language consists of three parts:

1. an abstract syntax, which specifies a class of annotation structures;
2. a formal semantics, describing the meaning of the annotation structures defined by the abstract syntax;
3. a concrete syntax, specifying a reference format for the physical representation of annotation structures defined by the abstract syntax.

Abstract and concrete syntax should moreover be related through the requirements that the concrete syntax is *complete* and *unambiguous* relative to the abstract syntax. These notions are defined as follows:

(8) a. **Completeness**: The concrete syntax defines a representation for every structure defined by the abstract syntax. (Possibly more than one, allowing alternative representations of the same abstract structure.)

b. **Unambiguity**: Every expression defined by the concrete syntax represents one and only one structure defined by the abstract syntax.

The representation format defined by a concrete syntax which has these two properties is called an *ideal representation format*. The property of 'completeness' means that there is a function R which to every structure α, defined by the abstract syntax, assigns a nonempty set $R(\alpha)$ of representations defined by the concrete syntax. Conversely, the property of 'unambiguity' means that there is a function R^{-1} which assigns to every expression e, defined by the concrete syntax, an annotation structure $R^{-1}(e)$ defined by the abstract syntax.

An important aspect of this design methodology is that the semantics of the annotation language is defined for the *abstract* syntax; given an expression e defined by the concrete syntax, its meaning is that of the annotation structure $R^{-1}(e)$. This ensures that any ideal representation format is convertible through a meaning-preserving mapping to any other ideal representation format.[2] In Ide & Bunt (2010), a mapping strategy is defined to convert from an abstract syntax to a representation in GrAF format (Ide & Suderman, 2007), and is illustrated with several annotation schemes, such as TimeML, PropBank, and FrameNet.[3] In addition to allowing for discourse annotation schemes to be represented uniformly across languages, domains, and genres, this may be useful to allow for effective combination of PDTB with GrAF renderings of PropBank and other annotations that have been done on the *WSJ*, including Penn Treebank (PTB) syntactic annotations.

[2]See Bunt (2010; 2011) for formal definitions and proofs.

[3]GrAF may be considered as a pivot format into which well-formed annotation schemes may be mapped, thus guaranteeing syntactic consistency and completeness for the purposes of comparison, merging, and transduction to other formats.

Taking the PDTB annotation scheme as the starting point for defining an ISO standard for the annotation of discourse relations, the first steps in this direction are to translate the PDTB scheme into an abstract syntax form, and to specify a concrete XML syntax for representing the annotation structures. This is the subject of the next two subsections.

4.2. Abstract syntax

The abstract syntax of DRelML consists of: (a) a specification of the elements from which annotation structures are built up, a 'conceptual inventory', and (b) a specification of the possible ways of combining these elements.

a. Conceptual inventory

The conceptual inventory of DRelML consists of a number of disjoint sets whose elements provide the ingredients for building annotation structures for discourse relations. Since a discourse relation in the PDTB is always a binary relation, with two arguments, the ingredients we need are those for identifying a discourse relation and its two arguments, including their attributions.

Since annotations add linguistic information to certain regions of primary data, such as particular stretches of text or speech, the annotation of a discourse relation includes the identification of the regions of primary data corresponding to the arguments of the relation, and in the case of an explicit discourse relation (expressed by a connective or by another type of expressions) also the region where the relation is expressed. In stand-off format, this is done through pointers to the primary data or to elements at another layer of annotation where the regions of primary data are identified. Following ISO practice, we will use the term 'markable' to refer to the entities that anchor an annotation directly or indirectly in the primary data. The conceptual inventory therefore also includes a set of markables. Altogether, the conceptual inventory therefore consists of the following sets:

1. *DR*, a finite set of discourse relations, $R_1, R_2, \ldots, R_n$. The hierarchical organization of the PDTB set of discourse relations, with lower tiers expressing more fine-grained meanings, is as such not part of the conceptual inventory, but follows from the definitions of each of these relations (cf. (Miltsakaki et al., 2008)).
2. *EntRel*, a singleton set containing a coherence relation, expressing that two sentences are related due to semantic relations between entities mentioned in the two sentences, such as coreference.
3. *MA*, a finite set of markables to which discourse relations information can be attached.
4. Four finite sets of features of attributions – *source*, *type*, *polarity*, and *determinacy*: *AtS* (attribution source), *AtT* (attribution type), *AtP* with two values for scopal polarity, and *AtD* with two values for the determinacy of an attribution.

5. *AOType*, a finite set of abstract object semantic types, $ao_1, ao_2, \ldots, ao_n$. Compared to the PDTB this is a new annotation category that we have introduced in order to make room for specifying semantic information about arguments, if desired. As with the discourse relations, inheritance relations hold between object types; these are based on the hierarchical classification in Asher (1993).

b. Annotation structures

An annotation structure is a set of *entity structures* and *link structures*. An entity structure contains semantic information about a region of primary data, as identified by markables; a link structure describes a semantic relation between the contents of two such regions. DRelML annotations can refer to six kinds of markables, described below.

Entity structures: An entity structure is one of the following structures:

a. *Explicit Attribution Entity Structure*, which is a pair $\langle m, a \rangle$ consisting of a markable m and an an *'Attribution Information Structure'* a, which is one of the following structures:

- $\langle as \rangle$;
- $\langle as, at \rangle$;
- $\langle as, ap, ad \rangle$;
- $\langle as, at, ap, ad \rangle$,

where $m \in MA$, $as \in AtS$, $at \in AtT$, $ap \in AtP$, and $ad \in AtD$,

The different possible structures capture the fact that, if attribution is annotated for discourse relations and their arguments, the scheme is still flexible with respect to what exactly is annotated. Minimally, only the text span signaling the attribution is marked and a source. In the other structures, one or more additional semantic features are also annotated, including the semantic type, polarity and determinacy of the attribution.

As the name suggests, Explicit Attribution Entity Structures will be used to annotate explicit attributions, while Attribution Information Structures will be used for annotating implicit ones. For short, we will also use the term *Attribution Structure* to designate either an Explicit Attribution Entity Structure or an Attribution Information Structure.

b. *Explicit Relation Entity Structure*, which is one of the following structures:

1. $\langle m, r \rangle$; $\langle m, r, a \rangle$; $\langle m, r, m_{hd}, m_{mod} \rangle$; $\langle m, r, a, m_{hd}, m_{mod} \rangle$;
2. $\langle m, r_1, r_2 \rangle$; $\langle m, r_1, r_2, a \rangle$; $\langle m, r_1, r_2, m_{hd}, m_{mod} \rangle$; $\langle m, r_1, r_2, a, m_{hd}, m_{mod} \rangle$.

where m is a markable, r, r_1, $r_2 \in DR$ are discourse relations, a is an Attribution Structure, and m_{hd} and m_{mod} are markables identifying the head and modifier(s) of a discourse connective, respectively.

The phenomenon that discourse connectives can have multiple senses is captured by the possible structures in (ii), with two senses (r_1 and r_2). Only up to two senses are allowed. Note that all structures occur with and without an Attribution Structure and with and without a connective head and modifier specification. This means that these elements are optional.

c. *Argument Entity Structure*, which is one of the following structures:

$\langle m \rangle$; $\langle m, a \rangle$; $\langle m, a, ao \rangle$

where m is a markable, a is an *Attribution Structure*, and $ao \in AOType$ is an abstract object type. Three different structures are defined, in order to allow the argument to be annotated with an attribution and/or with an abstract object type, without making any of them obligatory.

Link structures: A link structure is one of the following:

- An *Explicit Discourse Relation Structure*, which is a triple $\langle Arg1, Arg2, R \rangle$, consisting of two *Argument Entity Structures*, *Arg1* and *Arg2*, and an *Explicit Relation Entity Structure*, *R*.
- An *Implicit Discourse Relation Structure* is one of the following structures:

  i. $\langle Arg1, Arg2, r \rangle$; $\langle Arg1, Arg2, r, a \rangle$,

  ii. $\langle Arg1, Arg2, r_1, r_2 \rangle$; $\langle Arg1, Arg2, r_1, r_2, a \rangle$

  where *Arg1* and *Arg2* are *Argument Entity Structures*, $r, r_1, r_2 \in DR$ are discourse relations, and a is an Attribution Structure.

  As in the case of an Explicit Relation Entity Structure, the two variants in ii. capture the phenomenon that two sentences may be semantically related by more than one discourse relation (maximally two); the occurrence of variants with and without an Attribution Structure means that attributions of arguments are treated as optional.
- An *Entity Relation Structure*, $\langle Arg1, Arg2, E \rangle$ consisting of the entity-based coherence relation Et and two arguments *Arg1, Arg2*, which are either just a markable $\langle m \rangle$ or a pair $\langle m, a \rangle$ where $ao \in AOType$ is an abstract object type.

4.3. Concrete syntax

Given the abstract syntax defined above, an XML-based concrete syntax of DRelML is defined by applying the notion of an ideal representation format, defined above. As described in Bunt (2010), an ideal XML-based representation format can be defined systematically by designing XML elements and attributes to correspond to object types and their properties. For DRelML this means the definition of the following representation structures.

1. For each type of entity structure, defined by the abstract syntax, define an XML element with the following attributes:

(a) one for each component of the entity structure;

(b) the attribute `xml:id`, whose value is a unique identifier of the entity structure;

(c) the attribute `target`, whose value refers to a markable.

2. For each type of link structure, define an XML element with attributes whose values represent a relation and its arguments.

The notion of an ideal representation forma allows the introduction of extra attributes and values in the concrete syntax, because of their convenience for annotators, or their usefulness for certain annotation purposes, as long as these additional components do not interfere with the requirements of completeness and umanbiguity.

Concretely, in order to be maximally compatible with the PDTB, attributes/values are introduced for representing supplementary argument regions, inserted connectives for implicit discourse relations, and the distinction between explicit discourse relations expressed by connectives and those expressed by other means ('AltLex'). Altogether, this leads to the following concrete syntax definition:

Entity structure representations

1. an XML element called `dRelArgument`, wich has the following attributes:
   - `xml:id`, whose value specifies a unique identifier;
   - `target`, whose value identifies a markable;
   - `attribution`, whose value represents an explicit or implicit attribution (*optional*);
   - `aoType`, whose value specifies the abstract object type denoted by the markable (*optional*);
   - `supplRegion`, whose value represents a supplementary markable (*optional*).
2. an XML element called `explDRel`, which has the following attributes:
   - `xml:id`, whose value specifies a unique identifier;
   - `target`, whose value represents a relational markable;
   - `synType`, whose value indicates whether an explicit discourse relation is expressed by a connective (the value `connective`) or by some other kind of expression (the value `altLex`) (*optional*);
   - `headConn`, whose value represents the lexical head of a discourse relation expressed by a connective (*optional*);
   - `modConn`, whose value represents the modifier, if present, of a discourse relation expressed by a connective (*optional*);
   - `attribution`, whose value represents an explicit or implicit attribution (*optional*);
   - `discRel`, whose value names a discourse relation.
3. an XML element called `implDRel`, which has the following attributes:
   - `xml:id`, whose value specifies a unique identifier;
   - `discRel`, whose value names a discourse relation;
   - `disConn`, whose value represents a connective, inserted for an implicit discourse relation (*optional*).
4. An XML element called `explAttribution`, which has the following attributes:
   - `xml:id`, whose value specifies a unique identifier;
   - `target` whose value identifies a markable;
   - `atSource`, whose value represents the agent or other kind of source to whom a discourse relation or an argument of a relation is attributed;
   - `atType`, whose value represents the kind of attribution (*optional*; for the PDTB, the possible values are *PAtt, Ftv, Ctrl, Undef*));
   - `atPolarity`, whose value represents the scopal polarity, possibly associated with a negated discourse relation (*optional*);
   - `atDeterminacy`, whose value represents the determinacy of the attribution (*optional*).
5. An XML element called `implAttribution`, which has the same attributes as an `explAttribution`, except that it does not have a `target` attribute, being a non-consuming tag.

Link structure representations

- an element called `discourseRelation`, which has the following attributes:
  - `xml:id`, whose value specifies a unique identifier;
  - `arg1` and `arg2`, whose values are `dRelArgument` elements representing the arguments of the relation;
  - `rel1` and `rel2`, whose values are both either an `explDRel` or an `implDRel` element, representing the explicit or implicit discourse relations between the two arguments; `rel1` is obligatory; `rel2` is optional and used only when the two arguments are related by two discourse relations.
- an element called `entityRelation` which has two attributes: `arg1` and `arg2`, whose values refer to two `dRelArgument` elements, and the attribute `rel` which has the value `entityRel`;

5. Examples

(9) Example of the representation of a simple explicit discourse relation, with temporal connective *since*:

```
<dRelML>
<discourseRelation xml:id="dr1"
  arg1="#a1"
  arg2="#a2"
  rel="#er1"/>
<dRelArgument xml:id="a1"
  target="#m1"
  attribution="#at1"/>
<dRelArgument xml:id="a2"
  target="#m3"
  attribution="#at1"/>
<explRel xml:id="er1"
  target="#m2"
  discRel="succession"
  attribution="#at1"/>
<attributionInfo xml:id="at1"
  aSource="ot"/>
</dRelML>
```

(10) Example of the representation of a multifunctional discourse marker, with the connective *since* in temporal and causal interpretation:

```
<dRelML>
<discourseRelation xml:id="dr1"
  arg1="#a1"
  arg2="#a2"
  rel1="#er1"
  rel2="#er2"/>
<dRelArgument xml:id="a1"
  target="#m1"
  attribution="#at1"/>
<dRelArgument xml:id="a2"
  target="#m3"
  attribution="#at1"/>
<explRel xml:id="er1"
  target="#m2"
  discRel="succession"
  attribution="#at1"/>
<explRel xml:id="r2"
  target="#m2"
  discRel="reason"
  attribution="#at1"/>
<implAttribution xml:id="at1"
  aSource="ot"/>
</dRelML>
```

(11) An implicit simple discourse relation (conjunction), with different attribution sources of the two arguments:

```
<dRelML>
<discourseRelation xml:id="dr1"
  arg1="#a1"
  arg2="#a2"
  rel="#ir1"/>
<dRelArgument xml:id="a1"
  target="#m1"
  attribution="#at1"/>
<dRelArgument xml:id="a2"
  target="#m2"
  attribution="#at2"/>
<explAttribution xml:id="at1"
  target="#m3"
  aSource="ot"
  aType="comm"/>
<implAttribution xml:id="at2"
  aSource="wr"/>
<implRel xml:id="ir1"
  discRel="conjunction"
  attribution="#at1"/>
</dRelML>
```

(12) An implicit multiple discourse relation (conjunction and comparison):

```
<dRelML>
<discourseRelation xml:id="dr1"
  arg1="#a1"
  arg2="#a2"
  rel1="#ir1"
  rel2="#ir2"/>
<dRelArgument xml:id="a1"
  target="#m1"
  attribution="#at1"/>
<dRelArgument xml:id="a2"
  target="#m2"
  attribution="#at2"/>
<attributionInfo xml:id="at1"
  target="#m3"
  aSource="ot"
  aType="comm"/>
<attributionInfo xml:id="at2"
  aSource="wr"/>
<implRel xml:id="ir1"
  discRel="conjunction"
  attribution="#at1"/>
<implRel xml:id="ir2"
  discRel="comparison"
  attribution="#at1"/>
</dRelML>
```

6. Conclusions and perspectives

The exercise of creating an abstract syntax for the PDTB annotation scheme and rendering it in a graphic form shows the structure of the annotations clearly. The resulting concrete syntax is much more readable than the original format, and therefore errors and inconsistencies may be more readily identified. Furthermore, because it is rendered in XML, annotations can be validated against an XML schema (including validation that attribute values are among a list of allowable alternatives).

The abstract syntax also shows clearly an overall structure for a general-purpose standard for annotating discourse relations. We envision that any general-purpose discourse annotation scheme must allow for annotation based on all or any of several perspectives on elements of the task, such as semantic, interpersonal/intentional, and stylistic/textual, as identified in Hovy (1995). PDTB annotations are classified as "informational" (semantic, inter-propositional, ideational, pragmatic); the intentional and textual perspectives lie outside the scope of PDTB. PDTB's attribution types and the set of semantic classes, combined with those of other schemes, provide a base for a structured set of discourse annotation classes for the ISO

specification along the various axes of perspective, and at different levels of granularity.

Several topics for further work in developing an ISO standard for discourse relation annotation have emerged during the work reported in this paper. First, the approach underlying the PDTB has limited its scope to the annotation of relations between adjacent sentences. This limitation has been motivated by practical considerations regarding the work of human annotators. From a semantic point of view, however, both discourse relations within sentences and between non-adjacent sentences may be important. Second, the formal semantics of the abstract syntax still has to be worked out. Third, the establishment of sets of annotation concepts that are more broadly important than for the WSJ should deserve careful consideration, taking a range of languages, domains, and genres into account. This concerns in particular the set of discourse relations, and the sets of values used for the characterization of attributions (such as the set *Writer, Other, Arbitrary, Inherited* used in the PDTB). Explicit definitions of all the concepts, finally chosen as part of the standard, will have to be provided, and inserted in the ISOCat data registry.[4] Finally, the standard will not only have to define annotation and representation structures and concepts, but also examples and guidelines for their use in a range of practical situations.

7. References

Nicholas Asher (1993) *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Nicholas Asher and Alex Lascarides (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Harry Bunt (2010) A methodology for defining semantic annotation languages exploiting syntactic-semantic isomporphisms. In: Alex Chengyu Fang, Nancy Ide and Jonathan Webster (eds.) *Proceedings of ICGL 2010, Second International Conference on Global Interoperability for Language Resources*, Hong Kong, pp. 29-45.

Harry Bunt (2011) Defining languages for semantic annotation with an abstract syntax and a formal semantics. *Journal of Natural Language Engineering*, to appear.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum (2010) Towards an ISO standard for dialogue act annotation. In *Proceedings of LREC 2010*.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, and David Traum (2010) ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of LREC 2012*.

Jeroen Geertzen and Harry Bunt (2006) Measuring annotator agreement in a complex hierrchical dialogue act scheme. in *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, Sidney, pp. 126-133.

[4] See e.g. Kemps-Snijders, Windhouwer and Wright (2010) and `http://www.isocat.org`.

Barbara J. Grosz and Candace L. Sidner (1986) Attention, intentions, and the structure of discourse. In *Computational Linguistics* Vol. 12, n.3. pages 175-204.

Jerry Hobbs (1985) *On the coherence and structure of discourse*. Technical Report. Stanford University.

Ed Hovy (1995) The Multifuncitonality of Discourse Markers. in *Proceedings of the Workshop on Discourse Markers*, Egmond-aan-Zee, The Netherlands.

Nancy Ide and Harry Bunt (2010). Anatomy of Annotation Schemes: Mappings to GrAF. in *Proceedings of LAW-IV: the Fourth Linguistic Annotation Workshop*, Uppsala, pp. 115 -124.

Nancy Ide, Rashmi Prasad, and Aravind Joshi (2011). Towards Interoperability for the Penn Discourse Treebank, In *Proceedings of the 6th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-6)*, Oxford, U.K.

Nancy Ide and Laurent Romary (2004). International Standard for a Linguistic Annotation Framework. *Natural Language Engineering*, 10:211 - 225.

Nancy Ide and Keith Suderman (2007). GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the First Linguistic Annotation Workshop LAW-I*, held in conjunction with ACL 2007. Prague, pp. 1-8.

ISO (2006) ISO 24612:2006 Language resource management: Feature structures, Part 1: Feature structure representation. ISO, Geneva.

ISO (2010) ISO 24612:2010 Language resource management: Linguistic annotation framework (LAF), ISO, Geneva.

ISO (2010) ISO DIS 24617-2 Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO, Geneva.

Andrew Kehler (2002) *Coherence, Reference, and the Theory of Grammar*. CSLI Publications, Stanford.

Marc Kemps-Snijders, Menzo Windhouwer, and Sue Ellen Wright (2010) Standardizing Data Cetegories in ISOcat: Implementing Group Work for Thematic Domains. In *Proceedings Post-conference workshop at TKE 2010 Conference on Terminology and Knowledge Engineering*, Dublin.

Alan Lee, Rashmi Prasad, Aravind Joshi and Bonnie Webber (2008) Departures from tree structures in discourse: Shared arguments in the Penn Discourse Treebank. In *Proceedings of the Constraints in Discourse III Workshop*. Potsdam, Germany.

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

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz (1993) Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics* Vol.19, n.2. pages 313-330.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber (2004) Annotating discourse connectives and theoir arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*. Boston, MA. Pages 9-16.

Eleni Miltsakaki, Livio Robaldi, Alan Lee and Aravind Joshi (2008) Sense Annotation in the Penn Discourse

Treebank. In *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science* Vol. 4919. Pages 275-286.

Livia Polanyi (1987) *The Linguistic Discourse Model: Towards a formal theory of discourse structure*. Technical Report. Bolt Beranek and Newman, Inc.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi and Bonnie Webber (2007). Attribution and its Annotation in the Penn Discourse TreeBank. In *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse* Vol.47, n.2. pages 43-64.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

The PDTB Group (2008) *The Penn Discourse Treebank 2.0. Annotation Manual*. IRCS Technical Report IRCS-08-0.1 Institute for Research in Cognitive Science, University of Pennsylvania. Philadelphia, PA.

The Text Encoding Initiative (2007) P5 Guidelines for Electronic Text Encoding and Interchange, edited by Syd Bauman and Lou Burnard. The Text Encoding Initiative, Charlottesville, Virginia. Available at `http://www.tei-c.org/Guidelines/P5/`.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott (2003). In *Computational Linguistics* Vol.29, n.4. pages 545-587.

Florian Wolf and Edward Gibson (2005) Representing discourse coherence: A corpus-based study. In *Computational Linguistics* Vol.31, n.2.