Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation

Christian Chiarcos
Goethe-University Frankfurt am Main, Germany
chiarcos@informatik.uni-frankfurt.de

Abstract

This paper describes the extension of the Ontologies of Linguistic Annotation (OLiA) with respect to discourse features. The OLiA ontologies provide a a terminology repository that can be employed to facilitate the conceptual (semantic) interoperability of annotations of discourse phenomena as found in the most important corpora available to the community, including OntoNotes, the RST Discourse Treebank and the Penn Discourse Treebank. Along with selected schemes for information structure and coreference, discourse relations are discussed with special emphasis on the Penn Discourse Treebank and the RST Discourse Treebank. For an example contained in the intersection of both corpora, I show how ontologies can be employed to generalize over divergent annotation schemes.

Keywords: discourse annotation, OLiA ontologies, discourse structure, information structure, coreference, RSTDTB, PDTB

1. Background and motivation

Discourse annotations are often seen as an aspect of higher-level semantics, and naturally, possible bridges between semantic resources and technologies and discourse phenomena have been explored, e.g., in the application of WordNet to identify bridging relations (in coreference annotation) and in the application of FrameNet to identify certain discourse relations (in discourse structural annotation). This paper explores another way how both fields can be brought closer together: Using Semantic Web formalisms to formalize linguistic concepts applied in discourse annotations facilitates

interoperable representation Semantic Web standards such as the Resource Description Framework (RDF) are successfully applied to encode the output of NLP tools and diverse NLP resources in a well-defined interoperable format. Discourse annotations represent a natural continuation of these efforts.

development of shared terminologies RDF enforces the use of globally unambiguous IDs in the web of data: URI references can thus be used to link, or point to different, community-maintained vocabularies as (Linguistic) Linked Open Data.

terminology management Terminology repositories can be used to formalize the relation between different *labels* used in the existing multitude of independently developed discourse annotation schemes. For terminology maintenance, Semantic Web standards like the Web Ontology Language (OWL) can be employed.

axiomatization OWL2/DL provides a vehicle for the axiomatization of linguistic annotations, thereby formalizing classical decompositional approaches to, e.g., the study of discourse relations (Sanders, Ted and Spooren, Wilbert and Noordman, Leo, 1992).

Shared terminology repositories using Semantic Web formalisms may thus represent a suitable device to harmonize annotation terminology. This paper describes the development of such a resource specifically directed to annotations of discourse phenomena, i.e., discourse structure (subordinating or coordinating), coherence relations (semantic relations between individual utterances), coreference (anaphora, bridging), information status (given-new) and information structure (topic-focus).[1] In recent years, several approaches to establish conceptual (semantic) interoperability between linguistic annotations produced by different tools, according to different annotation schemes, or for different languages, have built on the creation and use of centralized terminology repositories of this type (Ide and Romary, 2004; Farrar and Langendoen, 2003). Yet, to my best knowledge, the approach described in this paper is the first approach to model annotation terminology specifically for discourse *annotations as used in major corpora available to the community*, whereas earlier approaches focused on developing more abstract upper models for discourse annotation (Goecke et al., 2005; Pareja-Lora and Aguado de Cea, 2010b; Pareja-Lora, 2012) or were restricted to a particular domain and/or theory (Ciccarese et al., 2008; Bärenfänger et al., 2008). It extends an existing modular architecture that has been previously applied for morphological, morphosyntactic and syntactic annotations in about 70 European and non-European languages and historical language stages, and whose capability to leverage heterogeneous annotations for different NLP applications (ensemble combination, NLP pipelines) and corpus exploration tools (cross-tagset queries and scripts) has been demonstrated before. Unlike earlier approaches on ontological models of discourse terminology for NLP and annotation, this modular architecture allows to explicitly provide ontological models of existing annotation schemes over which

[1] The broad field of multimodal communication is beyond the scope of this paper, as well as 'conversational' phenomena, e.g., speech acts, or dialog annotations that are covered by the Semantic Annotation Framework, Part 2: Dialogue acts (ISO 24617-2:2012) which may be directly adopted to extend the OLiA Reference Model. Related ISO standards for semantic relations in discourse (SemAF-DRel, ISO/WD 24617-8) and discourse structure (SemAF-DS, ISO/DTS 24617-5) are still under development (http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=297592) but may be integrated in the architecture upon publication.

then a generalization is developed. Related research has focused exclusively on either modeling an upper model for discourse terminology (Goecke et al., 2005; Ciccarese et al., 2008; Pareja-Lora and Aguado de Cea, 2010b; Pareja-Lora, 2012) or a specific theory, e.g., (Bärenfänger et al., 2008). The modular architecture applied here allows to leverage these ontologies and existing annotation schemes through Linking Models (see below).

This approach aims at establishing interoperability between existing corpora or on-going annotation initiatives in the computational linguistics, NLP and corpus linguistics communities by the explicit application of Linked Data principles for multiple, distributed, and modular ontologies. For coordination between these, an intermediate level of representation is introduced that mediates between several existing terminology repositories and multiple different annotation schemes.

The resulting modular architecture of OWL2/DL ontologies comprises the following components:

(i) **Annotation schemes** and **existing terminology repositories** are formalized as independent, self-contained ontologies (termed ANNOTATION MODELS and EXTERNAL REFERENCE MODELS, respectively).

(ii) Between both, an **intermediate** REFERENCE MODEL provides definitions which are are both derived from existing terminology repositories and harmonized with definitions found in the individual annotation schemes

(iii) Annotation Models and terminology repositories are **indirectly linked** through `subPropertyOf` and `subClassOf` ($\sqsubseteq$) descriptions that map annotation-specific terms to the Reference Model and Reference Model terms to terminology repositories (LINKING MODELS).

Introducing the OLiA Reference Model as intermediate representation reduces the number of mappings necessary to link every Annotation Model with every External Reference Model. Separating Annotation Model (resp. External Reference Model) and Linking Model for a particular annotation scheme establishes a clear separation between original documentation (in the Annotation Model) and its interpretation in terms of the Reference Model (in the Linking Model), so that the trustworthiness of both information sources can be assessed independently.

This design is well-established in the Ontologies of Linguistic Annotation – briefly, OLiA ontologies (Chiarcos, 2008) –, whose extension to discourse described in this paper. Previous applications of the OLiA ontologies include annotation interoperability for documentation and corpus querying, as well as NLP pipelines. It has been demonstrated in earlier research that the ontological representation of linguistic annotations allows to abstract from tool- or tagset-specific string representations of annotations, and that thus, information can be retrieved and integrated in a resource-independent fashion. In this way, the OLiA ontologies contribute to the establishment of conceptual (semantic) interoperability between tools and resources.

It is important to note in this context that the modeling of distributed ontologies by means of Linked Data principles extends beyond the resources described in this paper, but rather, puts them in direct relation with other terminology repositories.

2. Modeling discourse features

The OLiA architecture follows the principle to formalize linguistic phenomena *for annotation*, i.e., to create an intermediate level of representation for existing annotation schemes and existing terminology repositories. In the realm of discourse, however, no commonly accepted terminology repository is publicly available at the moment (cf. Sect. 5.), but only initial attempts to develop such repositories (Goecke et al., 2005; Bärenfänger et al., 2008; Rizzo and Troncy, 2011; Pareja-Lora and Aguado de Cea, 2010b; Pareja-Lora, 2012).

These ontologies have been consulted for the development of the discourse profile of the OLiA Reference Model, and if available, they are also linked to the Reference Model as External Reference Models. So far, however, this is the case only for the NERD ontology (Rizzo and Troncy, 2011) whose most important component is a taxonomy of lexical-semantic categories for named entities (hence, not directly discourse-related), for the SemDoc ontology of RST relations by Bärenfänger et al. (2008) and the minimalistic Grounded Annotation Framework.[2] The other ontologies are partially available to the authors, but not publicly released. Another important resource were, of course, the annotation schemes considered here, in particular, annotation schemes for coreference, information structure, discourse relations and discourse structure.

2.1. Top-level categories

The top-level category of the OLiA Reference Model is `LinguisticConcept`. For most levels of description, it is possible to distinguish between structural entities that are subject to annotation (e.g., `MorphosyntacticCategory` for parts-of-speech, or `SyntacticRelation` for edges in syntax annotation), and features that are assigned to these entities (e.g., `MorphosyntacticFeature` for inflectional morphology, or `SyntacticFeature` for edge labels in syntax annotation). Between entities and features hold the corresponding `hasFeature` relations (e.g., `hasAspect`, `hasSyntacticRole`). Every `LinguisticConcept` that can occur as object of a `hasFeature` predicate is defined as a subclass of `Feature`.

For discourse annotation, the following `LinguisticConcept`s were introduced: `DiscourseCategory` for non-relational structures and entities in discourse, `DiscourseRelation` for relations between `DiscourseCategory`s, and `DiscourseFeature` for annotations assigned to `DiscourseCategory`s and `DiscourseRelation`s.

[2] `http://groundedannotationframework.org/gaf-ontology/`, providing the concepts `Mention` (for markables) and `Instance` (for discourse entities) and the `denotes` property to link them.

2.2. Coreference and bridging

For coreference, the OLiA ontologies currently comprise Annotation Models for five different annotation schemes: the annotation scheme of the OntoNotes corpus (Hovy et al., 2006), the Potsdam Coreference Scheme (Krasavina and Chiarcos, 2007), applied to German and English newspaper corpora, the annotation scheme of the German TüBa-D/Z corpus (Naumann, 2007), the MATE-based (Poesio, 2004) annotation scheme of the English ARRAU corpus (Poesio and Artstein, 2008), and the annotation scheme for coreference and information structure applied to the German DIRNDL corpus (Riester et al., 2010).

The primary data structure of coreference annotation are *markable* (referring expressions) and *relations* (connecting markables that denote the same discourse entity). Bridging – or, textual inference – is closely related to coreference, but a bridging relation between two markables indicates that the discourse referents denoted by both markables are connected by a relation other than identity.

Accordingly, the `DiscourseCategory` `DiscourseEntity` was introduced to represent possible source and target elements of anaphoric and bridging relations. `DiscourseEntity` is defined with respect to the common ground established between hearer and speaker: A `DiscourseEntity` is any conceptual entity that can be introduced in the common ground by linguistic means, or that can be referred to by anaphoric means (e.g., a pronominal or definite description) if it is established in the common ground.

An important subclass of `DiscourseEntity` is `DiscourseReferent`. A `DiscourseReferent` is an object of conception, e.g., an entity in the real world, whose prototypical linguistic realization is by means of a nominal (or pronominal) expression, e.g., a name or a definite NP.

The definition of `DiscourseEntity` given above does, however, also cover utterances (resp. the propositions expressed by these utterances, and the states and events addressed by this proposition); (parts of) utterances can be referred to by pronouns: *When I first found out* $[$*I had diabetes*$]_i$ *I denied* $[it]_i$. Adopting a theory-neutral term established in the realm of discourse structure (Sect. 2.4.), such `DiscourseEntity`s are referred to as `DiscourseSegment` here.

As far as relations are concerned, the concept `EntityBasedRelation` $\sqsubseteq$ `DiscourseRelation` was established, with two subconcepts `AnaphoricRelation` and `BridgingRelation`.

Whereas coreference is actually an identity relation, there exists a multitude of taxonomies suggested for bridging relations between discourse referents. These subtypes of bridging are, however, not modeled as different relations, but as features that can be assigned to a `BridgingRelation`. The corresponding concept is `BridgingType` $\sqsubseteq$ `DiscourseFeature`.

2.3. Information structure and information status

At the moment, only few corpora with information structure annotations are available, and only two OLiA Annotation Models have been developed so far: One annotation model for the annotations of the German DIRNDL corpus (Riester et al., 2010), and one annotation model for the guidelines of Dipper et al. (2007), a set of guidelines applied to typological data collections for various languages (Skopeteas and Fanselow, 2009), and several corpora of German and its historical stages (Ritz et al., 2008).

Information structure is concerned with the structuring of utterances with respect to the type of information they convey, in particular the concepts of *topic* (what the utterance is about) and *focus* (what is said about the topic). Both `Focus` and `Topic` are modeled as subconcepts of `InformationStructuralEntity`, a subconcept of `DiscourseCategory`.

Information structure is closely related with information status, and it is concerned with the assessment to what extent a discourse referent is currently accessible (given) to the hearer, resp., present (and salient) in the common ground. Particularly salient referents are, for example, assumed to be more likely topics than non-salient referents. Information status is thus a property that can be assigned to a `DiscourseEntity`, and it is modeled as `InformationStatus` $\sqsubseteq$ `DiscourseFeature`.

2.4. Discourse structure and discourse relations

Theories of discourse structure address three main aspects, i.e., (1) **discourse structure** ('constituents' of discourse and their structure), (2) **discourse relations** (relations between utterances), and (3) **accessibility domains** (how relations and structure influence the realization and interpretation of utterances, e.g., through constraints on anaphora or information structure). Most theories aim to combine these aspects, or emphasize one of them. In terms of annotated corpora, the most important theories of discourse structure are the Rhetorical Structure Theory (RST), and discourse relations of the Penn Discourse Treebank (PDTB).

RST (Mann, W. and Thompson, S., 1988) defines discourse structure as a tree, where discourse segments are connected by subordinating (mononuclear) or coordinating (multinuclear) relations. Further, relations between discourse segments are distinguished with respect to their meaning or function, e.g., one discourse segment can represent express the cause, the justification, or just background information for the information conveyed by another discourse segment. Further, discourse structure interacts with anaphora (Cristea et al., 1998). Several corpora annotated with RST are available, the most important being the RST Discourse Treebank (Carlson et al., 2003, RSTDTB) for which an OLiA Annotation Model has been developed as described below.

Criticism on RST and related approaches (Wolf and Gibson, 2005) emphasizes practical and conceptual problems that the enforcement of discourse-structural constraints on discourse relations imposes: RST requires that discourse relations can only hold between discourse segments that are coordinated or subordinated and adjacent, but at least some discourse relations seem to be independent from discourse structural constraints (Webber et al., 2003), and enforcing these constraints on such discourse relations may lead to problems in inter-annotator agreement and reproducibility, an observation already made by Mann and Thompson

(1988).

Consequently, researchers have begun to disentangle discourse structure and discourse relations, and to develop annotation schemes that focus on discourse relations alone. In these schemes, annotating hierarchical discourse structure is either discouraged, e.g., in the Penn Discourse Graph Bank (Wolf and Gibson, 2005, PDGB), or optional, e.g., in the Penn Discourse Treebank (Prasad et al., 2008, PDTB). The PDTB has been particularly influential, and its scheme as been applied to various languages, e.g., Turkish (Zeyrek and Webber, 2008), Hindi (Oza et al., 2009), Italian (Tonelli et al., 2010) and Chinese (Zhou and Xue, 2012).

The Reference Model generalizes over both RST and PDTB-style annotations by distinguishing `CoherenceRelation`[3] for the annotation of relation types between discourse segments and `DiscourseStructuralPattern` to capture the differentiation between coordinating and subordinating discourse relations. For discourse segments, the concept `DiscourseSegment` is used, that can be independently motivated for the annotation of coreference (Sect. 2.2.).

`CoherenceRelation` and `DiscourseStructuralPattern` are subconcepts of `DiscourseRelation`, but unlike a thesaurus or a classical annotation scheme, OWL does not enforce sibling concepts to be disjoint. It is therefore possible to assign the same relation in a corpus a `CoherenceRelation` and a `DiscourseStructuralPattern`. Actually, (most) PDTB relations are defined as subconcepts of `CoherenceRelation`, whereas RST relations are defined by the intersection between a `CoherenceRelation` and a `DiscourseStructuralPattern`. In this way, the same inventory of coherence relations can be used for RST- and PTDB-annotated corpora (Fig. 1).

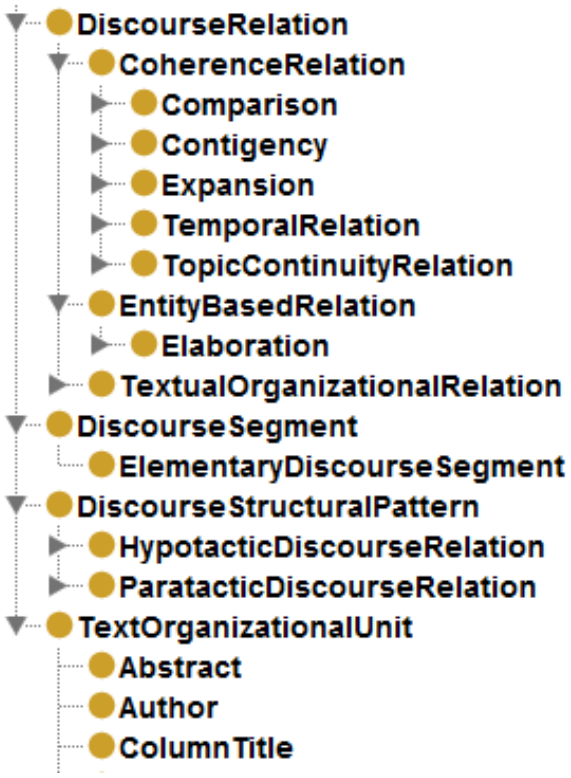

Figure 1: Selected discourse structure concepts in the Discourse Structure Model

3. A case study in interoperability

Using the OLiA ontologies, discourse annotations can be compared on a conceptual level. As an example consider a fragment of text (Fig. 2) that has been annotated for discourse structure in RSTDTB, PDTB and PDGB, and for coreference in OntoNotes. For brevity, we only discuss PDTB and RSTDTB annotations.

RSTDTB and PDTB have been developed with different theoretical backgrounds and different constraints on the annotations. For example, PDTB is missing ELABORATION, an important and generic class of discourse relations in RST. Moreover, PDTB eliminated the RST constraint that annotations eventually have to converge into a tree structure. Instead, discourse relations are annotated as relational structures. The figure shows conceptual discrepancies with respect to possible hierarchical structures: RSTDTB groups (5) together with (6) (and (7)) and then combines the resulting (5-7) segment with (4), whereas PDTB groups (4) and (5) together and assigns them (6) as an ARG2 argument.

Nevertheless, the *intuitions* about the semantics of the discourse relations seem to be consistent in both annotations, even though segmentation issues prevail. However, a segmentation-independent comparison of discourse relations seems feasible by comparing relations that *bridge the same gap*, e.g., between (5) and (6) (even though this is a relation between (5) and (6-7) in RST and between (4-5) and (6) in the PDTB). Table 1 gives the resulting pairings.

From these matches, the first seems to be a direct correspondence. Recognizing the relationship for the other two requires a slightly deeper understanding of the annotation schemes, but a causal aspect of EXPLANATION is relatively straight-forward, and alternative sets have been one classical way to formalize CONTRAST as a discourse relation. However, such an imprecise mapping is not necessarily a proper basis for an evaluation. Using ontologies, however, it is possible to represent more fine-grained nuances of meaning. If an annotation is known to correspond to an instance of concept A, we can automatically infer that it also is an instance of any B for any $A \sqsubseteq B$. An annotation a can thus be translated into a set of descriptions of the form $a \in A$. For evaluation purposes, this allows to quantify the number of shared descriptions between annotations of different sources – given a particular interpretation of annotation concepts as specified in the respective Linking Models.

Such an ontological interpretation can establish an improved level of conceptual interoperability: If two annotation schemes can be reduced to the same basic term in-

[3]The term 'coherence relation' follows (Hobbs, 1979) and (Kehler, 2002). The concept `DiscourseRelation` covers not only relations as considered in RST and PDTB, but also entity-based relations such as `AnaphoricRelation` and `NearIdentityRelation`.

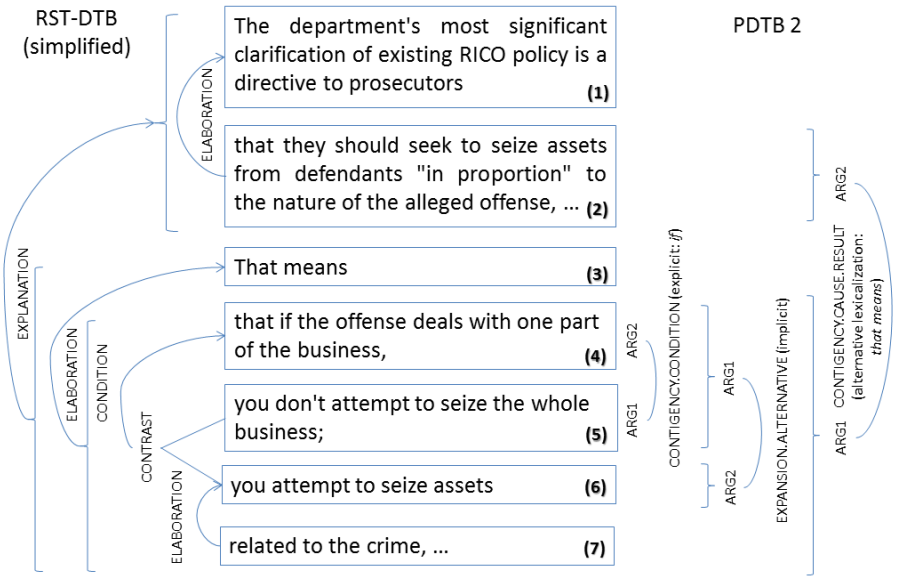

Figure 2: RSTDTB and PDTB annotations for *wsj_1365* (slightly simplified)

PDTB		RSTDTB	
(4)-(5)	CONTINGENCY.CONDITION.GENERAL	(4)-(5-7)	CONDITION
(2)-(4-7)	CONTINGENY.CAUSE.RESULT	(1-2)-(3-7)	EXPLANATION
(4-5)-(6)	EXPANSION.ALTERNATIVE.CHOSEN_ALTERNATIVE	(5)-(6-7)	CONTRAST

Table 1: Relation pairs from Fig. 2.

ventory, their information can be more easily compared. This perspective actually allows a direct comparison of relations from PDTB and RSTDTB, because the ontological model can be employed to disentangle semantic and structural aspects of discourse relations, the former within the `CoherenceRelation` branch, the latter within the `DiscourseStructuralPattern` sub-taxonomy, as shown in Tab. 2. The original annotations could have been less easily compared, because RST and PDTB Conditions overlap in some of their characteristics, but the string-level representation would not allow to represent these differences in meaning appropriately.

As such, the relations connecting (4) and (5) are described as `pdtb:GeneralCondition` and `rst:Condition`. The Reference Model provides a hierarchy that harmonizes PDGB, PDTB and RST relations. Its upper levels follow the PDTB structure, as it is the most general system of discourse relations *found in annotations* (there are equally justified alternative abstractions, but without annotations). `Condition` is a generalization over 'Condition' in PDGB, PDTB and RST. A dichothomy between `SemanticCondition` and `PragmaticCondition` reflects the differentiation between 'subject-matter' and 'presentational' relations in RST, thus the RST Condition is linked with the former concept. The PDTB `GeneralCondition` is a subconcept of `SemanticCondition` as it is defined in terms of truth values.

Integrating PDTB and RST information yields a gain in informativity about the discourse relation between (4) and (5): Only RST provides information about discourse structural hypotaxis, whereas the PDTB provides a more fine-grained classification of coherence relations (here).

Along these lines, all PDTB coherence relations can be compared to the RSTDTB discourse relations that link the corresponding segments, and regularities in that mapping can be further explored, and a richer representation can be achieved: Using the ontology, the original annotations can be represented in an interoperable way as a set of RDF triples that postulate relationships between the annotation and OLiA Reference Model concepts. Using standard set operations like intersection ($\sqcap$) and join ($\sqcup$), we can thus quantify their agreement ($\sqcap$) or integrate their information ($\sqcup$). And, of course, the modeling itself can be evaluated by measuring the relative number of shared triples between two annotation schemes for different Linking Models. However, this paper focuses on the *modeling* of terminologies. A study on the triple-based comparison and integration of PDTB and RSTDTB annotations is in preparation.

Even more interesting may be the ontologist's perspective, that is, by observing systematic correspondences, we may get a better, data-driven understanding and a comparable perspective on defining and correlative criteria of discourse relations. Potentially, this may lead to a revision of existing annotation schemes and theoretical models. For example, one can use the overlapping sections of the RSTDTB and the PDTB to bootstrap a more fine-grained set of discourse relations that either of these resources provided on its own, e.g., by subclassifying PDTB Contingency relations with respect to different discourse functions (as emphasized in the definition of Explanation).

PDTB Annotation & Linking	pdtb:contingency.condition.general a pdtb:GeneralCondition
	pdtb:GeneralCondition ⊑ olia:GeneralCondition_PDTB
OLiA Reference Model	olia:GeneralCondition_PDTB ⊑ **olia:SemanticCondition**
	olia:SemanticCondition ⊑ **olia:Condition**
	olia:Condition ⊑ **olia:Contingency**
	olia:Contingency ⊑ **olia:CoherenceRelation**
	olia:CoherenceRelation ⊑ **olia:DiscourseRelation**
	olia:ParatacticDiscourseRelation ⊑ olia:DiscourseStructuralPattern
	olia:DiscourseStructuralPattern ⊑ olia:DiscourseRelation
RST Linking & Annotation	rst:Condition ⊑ **olia:SemanticCondition** ⊓ olia:ParatacticDiscourseRelation
	rst:condition a rst:Condition

Table 2: Aligning RSTDTB and PDTB Conditions (shared superconcepts in **bold**)

Another important aspect of ontologies is that they provide a scalable level of granularity. It is thus possible to choose the appropriate level of detail for a particular purpose. Here, the upper levels of coherence relations follow the PDTB classification as this is the most coarse-grained classification of coherence relations applied to annotated data. At lower levels, PDTB, RST and PDGB relations have been aligned as shown above, and potentially, this allows to test different classification systems for discourse relations, thereby increasing the chance of finding a generalization that is easier to learn than any of the levels of granularity provided by the PDTB or RST relation hierarchies *as is*.

4. Results and discussion

This paper described the extension of the OLiA Reference Model with respect to discourse structure, discourse relations, information structure and coreference, three closely interrelated phenomena, for which a considerable bandwidth of annotated resources has become available in recent years: This is particularly true for coreference annotation which has been a flourishing line of research since the mid-1990s.

As for discourse structure, most corpora available follow either the RST tradition (Moser and Moore, 1995; Stede, M., 2004; van der Vliet et al., 2011), variants of the PDTB guidelines (Zeyrek and Webber, 2008; Oza et al., 2009; Tonelli et al., 2010; Huang and Chen, 2011; Zhou and Xue, 2012) or they try to amalgamate both (Buch-Kromann and Korzen, 2010). For these corpora, the concept inventory developed for the RST Discourse Treebank and the PDTB is applicable.[4]

Information structure is closely interrelated with both discourse structure (i.e., relations/transitions between utterances) and coreference (because of the impact of anaphora on information status and thus, the notions of topic and focus). The number of information-structurally annotated corpora available is, however, limited as compared to the other phenomena considered above. Aside from the annotation schemes mentioned above, further schemes for information structure have been developed, for example, for Czech (Hajič, 2005), English (Calhoun et al., 2005), Danish (Paggio, 2006) and a number of older Indo-European languages (Haug and Jøhndal, 2008).

Given the multitude of resources, this paper focused on a selection of representative schemes, without any claims with respect to exhaustivity or completeness, but with the goal to show how ontologies that generalize over different types of discourse annotations can be developed and to provide a publicly available (CC-BY) resource that may be employed as a nucleus for such an enterprise as part of the Linked Open Data cloud (see below). At the moment, the OLiA Reference Model for discourse[5] provides 263 concepts and 16 properties and are provided together with 4 Annotation Models for discourse structure, 4 Annotation Models for coreference and bridging, and 2 Annotation Models for information structure, 3886 triples in total.

For representative types of annotations and corpora, the approach described in this paper can thus be applied to ground different discourse-relevant terminologies in common specifications, thereby establishing conceptual interoperability between concepts used for discourse annotation. However, it should be emphasized that this level of interoperability only pertains to *labels* used in annotation, not for full structures. Although it is possible to query for `olia:Condition` and to retrieve annotations from both RSTDTB and PTDB, the *elements* that the annotated relations link are not comparable: PDTB `DiscourseSegments` are sentences or sentence fragments, whereas RSTDTB `DiscourseSegments` are nodes in a tree structure that may span large parts of the entire text.

Both representations can be partially mapped onto each other, but only through transformations, e.g., by applying Marcu's (Marcu, 1996) nuclearity principle to reduce RST `DiscourseSegments` to their nuclei, and hence to elementary discourse units that are more comparable to PDTB `DiscourseSegments`. Accordingly, full interoperability between RST and PDTB requires further developments in terms of *structural* interoperability. A possible generalization over both approaches may be seen in the conversion of discourse structural annotations to dependency DAGs

[4] Corpora developed in different theoretical frameworks include (S)DRT-based corpora (Asher, Nicholas and Lascarides, Alex, 2003), most noteably the Groningen Meaning Bank (Basile and Bos, 2011, GMB). So far, it has not been integrated, as its depth of discourse relation analysis is still limited: GMB 2.1.0 provides only three generalized discourse relations (continuation, contrast, parallel) for relations with explicit cues. Where other discourse markers are being used, their lemma is taken as relation label. (`http://gmb.let.rug.nl/data.php`)

[5] `http://purl.org/olia/discourse/olia_discourse.owl`

(Danlos, 2008). Such transformations (and similar considerations for markables in coreference annotation and information structure) may naturally complement the terminological harmonization represented in this paper, but are, at the moment, beyond its scope.

5. Perspectives and applications

The practical relevance of ontologies for NLP has been recognized since almost a decade, yet traditionally with a focus on natural language semantics (Cimiano et al., 2014). Unlike this, this paper aims not on formalizing the semantics of the text itself, but on developing formal vocabularies for the semantics of labels used in linguistic annotation.

The potential of such ontology-based specifications for linguistic annotations has been demonstrated by early studies like Pareja-Lora (2010a) and Chiarcos (2010). Both papers showed that the ontology-based semantic decomposition of morphosyntactic annotations generated by different NLP tools allowed to integrate their information more easily, regardless of the string representation of the original annotation. Using ontologes thus allows to develop ensemble combination architectures for NLP tools, but unlike existing string-based ensemble combination architectures, the resulting annotations are not only more robust, but also more fine-grained, as the tools are not required to use the same level of granularity, but rather, that tool-specific information may be preserved. Where subsequent processing steps require a specific annotation scheme, we may choose the appropriate level of detail to generate an appropriate string representation. More recently, the Semantic Web community developed NLP pipeline architectures that do no longer require conversion to string representations, but where subsequent processing modules directly communicate via ontological representations of their annotations.[6]

Task-specific discourse ontologies have been developed in the context of various NLP applications, e.g., for analyzing scientific discourse (Bärenfänger et al., 2008; Ciccarese et al., 2008), or named entity recognition (Rizzo and Troncy, 2011).

Within the OLiA approach, each of these ontologies can be integrated as an External Reference Model. Unlike these NLP-focused ontologies, that are specific to one particular annotation scheme, or that formalize only reference categories but not their relationship to annotations, OLiA focuses on the modeling of the mapping between annotation schemes and terminology repositories, i.e., the information about the *original* annotation is preserved, and the relationship between annotations and reference categories is defined in a declarative way such that any interpretation not originating from the external knowledge sources or the annotation documentation is represented transparently and reversible. On this basis, then, linguistic annotations from various sources can be interpreted in terms of several External Reference Models, and any algorithm developed for these External Reference Models can be applied to these annotations as well.

Discourse has also been suggested as a community-of-practice extension (Goecke et al., 2005) of the GOLD ontology (Farrar and Langendoen, 2003). Although never released to the public, it is modeled as an External Reference Model within the OLiA architecture. An extension of the OntoTag ontologies with respect to discourse and pragmatics has been presented by Pareja-Lora and Aguado de Cea (2010b; Pareja-Lora (2012). This ontology takes a different approach to the problem by formalizing discourse phenomena in a top-down fashion based on the theoretical literature. Also, it does not seem to be grounded in existing annotations or concrete NLP applications. Unfortunately, it is also not available to the public, otherwise, this would be a highly valuable external reference model.

Aside from NLP interoperability, another application of the OLiA ontologies can be seen in the development of corpus query systems where users can query for concepts from the OLiA Reference Model (or the External Reference Model of their choice, e.g., GOLD) instead of annotations. This naturally complements the development of generic data models for linguistic annotations, and corpus query engines developed on this basis, especially for multi-layer annotations (Rehm et al., 2007; Burchardt et al., 2008). As a result, it is possible to apply comparable corpus queries to different corpora, and even to evaluate multiple corpora at the same time. In this way, interoperable and portable corpus queries can be designed. These interoperable corpus queries, however, can be formulated in terms of either the OLiA Reference Model, or in terms of any terminology repository linked as an External Reference Model. From the perspective of a user, the difference is made clear through the use of different namespaces for the ontology concepts used in a query.

A modular architecture of independent ontologies as described here thus has the benefit of seamlessly integrating existing terminologies by the use of namespaces and RDF descriptions for the linking. Beginning with GOLD community-of-practice extensions, and the Typological Database System (TDS) (Saulwick et al., 2005) , this mechanism has been used to integrate diverse terminologies hosted by a particular organization. In recent years, however, researchers in our field are becoming increasingly aware of the potential of RDF and related standards to extend such links to physically separated collections of linguistic data, including not only terminology repositories, but also lexical-semantic resources and annotated corpora. Some of these activities are currently bundled in the development of the Linguistic Linked Open Data (LLOD) cloud.[7] In this context, the publication of the TDS ontology under an open license has been announced, and using the OLiA ontologies (available under a CC-BY license), the concept of External Reference Models as employed here can be reinterpreted as a linking with these resources , and thus provides a nucleus for the integration of terminology in the LLOD cloud.

For any discourse-annotated data available as Linked Data, e.g., an RDF output of Boxer (Augenstein et al., 2012), the semantic parser underlying the Groeningen Meaning Bank,[8] then, the OLiA ontologies already provide the nec-

[6] http://nlp2rdf.org

[7] http://linguistics.okfn.org/llod

[8] http://gmb.let.rug.nl/about.php

essary level of conceptual interoperability to combine and to integrate their linguistic annotations up to the discourse level, thereby allowing to run SPARQL queries against multiple discourse-relevant resources.

Acknowledgements

The research on discourse relations described in this paper was originally conducted during a PostDoc fellowship of the German Academic Exchange Service (DAAD) at the Information Sciences Institute ot the University of Southern California, and subsequently continued and extended at the Applied Computational Linguistics Lab at the Goethe-University Frankfurt, Germany, supported by the federal state of Hessen through the LOEWE cluster "Digital Humanities".

6. References

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

Augenstein, I., Padó, S., and Rudolph, S. (2012). Lodifier: Generating linked data from unstructured text. In *Proc. 9th Extended Semantic Web Conference (ESWC 2012)*, pages 210–224, Heraklion, Greece.

Basile, V. and Bos, J. (2011). Towards generating text from Discourse Representation Structures. In *Proc. 13th European Workshop on Natural Language Generation (ENLG 2011)*, Nancy, France, Sep.

Buch-Kromann, M. and Korzen, I. (2010). The unified annotation of syntax and discourse in the copenhagen dependency treebanks. In *Proc. 4th Linguistic Annotation Workshop (LAW-2010)*, pages 127–131, Uppsala, Sweden.

Burchardt, A., Padó, S., Spohr, D., Frank, A., and Heid, U. (2008). Formalising multi-layer corpora in OWL/DL. In *Proc. 3rd International Joint Conf on NLP (IJCNLP 2008)*, Hyderabad.

Bärenfänger, M., Hilbert, M., Lobin, H., and Lüngen, H. (2008). OWL ontologies as a resource for discourse parsing. *LDV-Forum*, 23(1):17–26.

Calhoun, S., Nissim, M., Steedman, M., and Brenier, J. (2005). A framework for annotating information structure in discourse. In *Proc. Frontiers in Corpus Annotations II*, pages 45–52.

Carlson, L., Marcu, D., and Okurowski, M. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In van Kuppevelt, J. and Smith, R., editors, *Current and New Directions in Discourse and Dialogue*, pages 85–112. Kluwer, Dordrecht.

Chiarcos, C. (2008). An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Chiarcos, C. (2010). Towards robust multi-tool tagging. An OWL/DL-based approach. In *Proc. 48th Annual Meeting of the Association for Computational Linguistics*, pages 659–670, Uppsala, Sweden, July.

Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., and Clark, T. (2008). The SWAN biomedical discourse ontology. *Journal of Biomedical Informatics*, 4:739–751.

Cimiano, P., Unger, C., and McCrae, J. (2014). *Ontology-Based Interpretation of Natural Language*. Morgan & Claypool, San Rafael, CA.

Cristea, D., Ide, N., and Romary, L. (1998). Veins Theory: A model of global discourse cohesion and coherence. In *Proc. ACL-98/COLING-98*, pages 281–285, Montréal, Canada.

Danlos, L. (2008). Strong generative capacity of RST, SDRT and discourse dependency DAGs. In Kühnlein, P. and Benz, A., editors, *Constraints in Discourse*. John Benjamins, Amsterdam; Phildelphia.

Dipper, S., Götze, M., and Skopeteas, S., editors. (2007). *Information Structure in Cross-Linguistic Corpora*. Interdisciplinary Studies on Information Structure (ISIS); 7. Universitätsverlag Potsdam, Potsdam, Germany.

Farrar, S. and Langendoen, D. T. (2003). A Linguistic Ontology for the Semantic Web. *GLOT International*, 7:97–100.

Goecke, D., Lüngen, H., Sasaki, F., Witt, A., and Farrar, S. (2005). GOLD and discourse. In *Proc. E-MELD Workshop on Morphosyntactic Annotation and Terminology*, Cambridge, Massachusetts.

Hajič, J. (2005). Complex corpus annotation: The Prague dependency treebank. In *Insight into the Slovak and Czech Corpus Linguistics*, pages 54–59. SAV, Bratislava, Slovakia.

Haug, D. and Jøhndal, M. (2008). Creating a parallel treebank of the old Indo-European Bible translations. In *Proc. Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008)*, pages 27–34, Marrakech, Morocco.

Hobbs, J. (1979). Coherence and coreference. *Cognitive Science*, 6:67–90.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proc. HLT-NAACL 2006*, pages 57–60, New York, June.

Huang, H. and Chen, H. (2011). Chinese discourse relation recognition. In *Proc. ICJNLP-2011*, pages 1442–1446, Chiang Mai, Thailand, Nov.

Ide, N. and Romary, L. (2004). A registry of standard data categories for linguistic annotation. In *Proc. LREC 2004*, pages 135–39, Lisboa, Portugal, May.

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

Krasavina, O. and Chiarcos, C. (2007). PoCoS - Potsdam Coreference Scheme. In *Proc. 1st Linguistic Annotation Workshop (LAW-2007)*, pages 156–163, Prague, Czech Republic, June.

Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory. *Text*, 8(3):243–281.

Marcu, D. (1996). Building up rhetorical structure trees. In *13th National Conference on Artificial Intelligence*, pages 1069–1074, Portland, Oregon.

Moser, M. and Moore, J. (1995). Investigating cue selection and placement in tutorial discourse. In *Proc. ACL-1995*, pages 130–135.

Naumann, K. (2007). Manual for the annotation of in-document referential relations. Technical report, Universität Tübingen.

Oza, U., Prasad, R., Kolachina, S., Sharma, D., and Joshi, A. (2009). The Hindi Discourse Relation Bank. In *Proc. 3rd Linguistic Annotation Workshop*, pages 158–161.

Paggio, P. (2006). Annotating information structure in a corpus of spoken Danish. In *Proc. LREC-2006*, pages 24–30, Genova.

Pareja-Lora, A. and Aguado de Cea, G. (2010a). Ontology-based interoperation of linguistic tools for an improved lemma annotation in Spanish. In *Proc. LREC 2010*, Valetta, Malta.

Pareja-Lora, A. and Aguado de Cea, G. (2010b). Modelling discourse-related terminology in OntoLingAnnot's ontologies. In *Workshop on Establishing and using ontologies as a basis for terminological and knowledge engineering resource*, Dublin, Ireland, August.

Pareja-Lora, A. (2012). OntoLingAnnot's Ontologies. In Chiarcos, C., Nordhoff, S., and Hellmann, S., editors, *Linked Data in Linguistics*. Springer, Heidelberg.

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In *Proc. LREC-2008*, Marrakech, Marocco.

Poesio, M. (2004). The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proc. SIGDIAL-2004*.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proc. LREC 2008*, Marrakech, Morocco.

Rehm, G., Eckart, R., and Chiarcos, C. (2007). An OWL- and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In *Proc. RANLP 2007*, Borovets, Bulgaria.

Riester, A., Lorenz, D., and Seemann, N. (2010). A recursive annotation scheme for referential information status. In *Proc. LREC-2010*, pages 717–722.

Ritz, J., Dipper, S., and Götze, M. (2008). Annotation of information structure. In *Proc. LREC-2008*.

Rizzo, G. and Troncy, R. (2011). NERD: Evaluating named entity recognition tools in the web of data. In *Workshop on Web Scale Knowledge Extraction (WEKEX 2011)*.

Sanders, Ted and Spooren, Wilbert and Noordman, Leo. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Saulwick, A., Windhouwer, M., Dimitriadis, A., and Goedemans, R. (2005). Distributed tasking in ontology mediated integration of typological databases for linguistic research. In *Proc. 17th Conf. on Advanced Information Systems Engineering (CAiSE'05)*, Porto.

Skopeteas, S. and Fanselow, G. (2009). Effects of givenness and constraints on free word order. In Zimmermann, M. and Féry, C., editors, *Information Structure. Theoretical, Typological, and Experimental Perspectives*. Oxford University Press, Oxford.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proc. ACL-2004 Workshop on Discourse Annotation*, Barcelona, Spain.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi, A. (2010). Annotation of discourse relations for conversational spoken dialogs. In *Proc. LREC 2010*, pages 2084–2090, Valletta, Malta.

van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., and Redeker, G. (2011). Building a discourse-annotated Dutch text corpus. In *Proc. Beyond Semantics*, pages 157–171, Göttingen, Germany, Feb.

Webber, B., Stone, M., Joshi, A., and Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence. *Computational Linguistics*, 31(2):249–287.

Zeyrek, D. and Webber, B. (2008). A discourse resource for Turkish. In *Proc. IJCNLP-2008*.

Zhou, Y. and Xue, N. (2012). PDTB-style discourse annotation of Chinese text. In *ACL-2012*, Jeju Island, Korea.