Originally published in: Ontologies in Text Technology: Approaches to Extract Semantic Knowledge from Structured Information. - Osnabrück: Institute of Cognitive Science, 2007. Pp. 1-5.

# Using OWL ontologies in discourse parsing

## – Extended abstract –

Maja Bärenfänger, Mirco Hilbert, Henning Lobin, Harald Lüngen

## 1 Introduction

In the project SemDok funded by the German Research Foundation DFG, a discourse parser for a complex type, i.e. scientific articles, is being developed. Discourse parsing (henceforth DP) according to Rhetorical Structure Theory (RST) ([1], [2]) deals with automatically assigning a text a tree structure in which discourse segments and rhetorical relations such as Concession between them are marked. For identifying the combinable segments, declarative rules are employed, which describe linguistic and structural cues and constraints about possible combinations by referring to different XML annotation layers of the input text, and external knowledge bases such as a discourse marker lexicon, a lexical-semantic ontology (later to be combined with a domain ontology), and an ontology of rhetorical relations. In our text-technological environment, the obvious choice of formalism to represent such ontologies is OWL ([3]). In this paper, we describe two OWL ontologies and how they are consulted from the discourse parser to solve certain tasks within DP. The first ontology is a taxononomy of rhetorical relations which was developed in the project. The second one is an OWL version of GermaNet, the model of which we designed together with our project partners.

## 2 Taxonomy of rhetorical relations

Already in the original conception of Rhetorical Structure Theory by Mann and Thompson [1], rhetorical relations were grouped into classes. On a top level, there were the two groups of *multinuclear* vs. *mononuclear* relations according to the structural criterion of nuclearity. The mononuclear relations were further subdivided into *presentational* vs. *subject-matter relations* cf. [1]. Lower-level subgroups such as *Evidence-and-Justify* were introduced as well. Hovy and Maier [4] suggested a merger of existing hierarchies of discourse relations into one comprehensive hierarchy consisting of 65 relation categories, 43 of which were relations at the base level. Their prediction was that application-specific extensions to this merged relation set would always consist in the refinement of

a relation category that was already in the hierarchy, i.e. the number of higher-level relation types would always stay the same. One purpose of developing a hierarchy of discourse relations is thus to point out similarities of different relation sets by showing how they can be mapped on each other or even merged, ultimately supporting the view that a universal set of relation types exists.

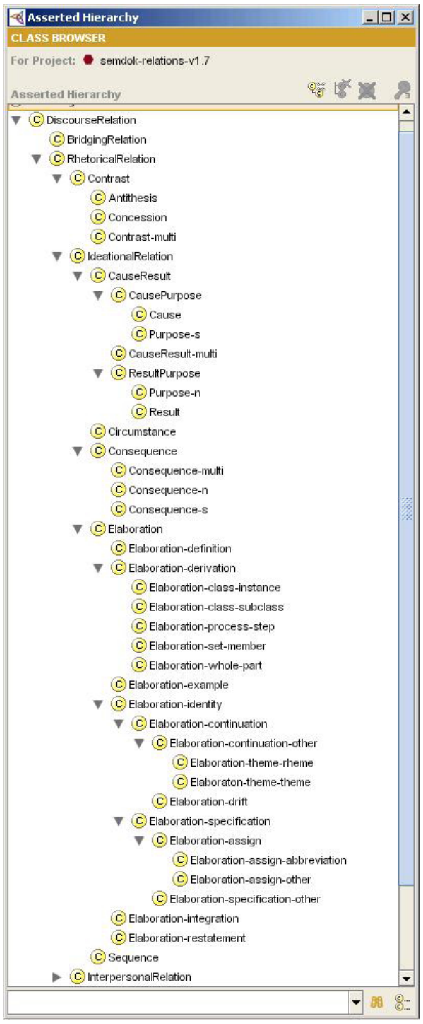


**Fig. 1.** Part of the RRSET ontology (screenshot of class browser in Protégé 3.1.1)

In the present project, we produced corpus annotations using the original RST relation set proposed in [1], and by an examination of these annotations

and an inspection of alternative relation sets proposed in the literature (notably [5] and [4]), we designed a relation hierarchy suitable for annotating the rhetorical structure of scientific journal articles in our explorative reading scenario [6]. It consists of 70 relation types, 44 of which are basic categories in the hierarchy.

Though it seems natural to model rhetorical *relations* as OWL *properties* (<owl:ObjectProperty>) as we proposed in the earlier publication [7], we finally refrained from doing so, because we also wanted to view the properties as classes to declare disjointness between certain rhetorical relation types and to encode properties of rhetorical relations that would be inherited by their subrelations. Within OWL DL, properties can be arranged in a hierarchy but cannot be declared classes at the same time ([3]).[1] Thus we modelled the rhetorical relations as OWL classes, which is not so devious if one considers that it is sometimes recommended to introduce a "relation class" for the encoding of an n-ary relation in OWL, cf. [8]. Subrelation-hood is then marked by the <rdfs:subclassOf> construct. The use of <rdfs:subclassOf> also enabled us to include further features in the formalisation of our hierarchy: We introduced heavily underspecified relation classes such as MONONUCLEARRELATION, and we cross-classified all relations along the two dimensions *nuclearity* and *metafunction*, giving rise to multiple inheritance. For example, SUPPORT is both the subclass of INTERPERSONALRELATION as well as of MONONUCLEARRELATION. We introduced further sub- or superrelations, when it was expedient according to our corpus analyses and with respect to our scenario, cf. [6]. The resulting hierarchy is shown in Fig. 1. This "RRSET ontology" is used to combine competing hypothesis during the parsing process as described in Sect. 4.

## 3 Using a GermaNet-based Ontology for the automatic assignment of ELABORATION

One of the most prominent RST relations in our corpus is ELABORATION - it is the second most frequent relation at all. Unlike other RST relations, ELABORATION is seldom signalled by syntactic or lexical discourse markers. To tackle its automatic identification and annotation, we examined instances of ELABORATION in our corpus and reviewed the treatment of ELABORATION in previous approaches to discourse analysis (e.g. [5], [4]). This led us to distinguish the different subtypes of ELABORATION relations which can be seen in the taxonomy of rhetorical relations in Fig. 1.

The subtaxonomy of ELABORATION relations organises the subcases that can trigger different types of rhetorical links between text modules of scientific articles in our explorative reading scenario. Each subrelation has its own definition and is associated with a different set of discourse markers and linguistic or structural cues that signal it. ELABORATION-DEFINITION, for example, can be determined by cues from the logical document structure (e.g. <doc:glosslist>),

[1] Since most OWL reasoners and inference tools apply to the sublanguage OWL DL, we encode our ontologies within OWL DL.

ELABORATION-EXAMPLE is often signalled by the lexical discourse markers"z.B.", "Beispiel", or "beispielsweise"), whereas the subtypes of ELABORATION-SPECIFICATION are induced by syntactic and punctuational discourse markers (e.g. a non-sentential phrase within parentheses).

For the most frequent subtypes of ELABORATION, an OWL version of the lexical-semantic net GermaNet ([9]) shall be consulted: ELABORATION-DERIVATION is established by the presence of conceptual relations like hyperonymy/ hyponymy, holonymy or meronymy between the central discourse entities (themes) of two discourse segments, and lexical relations like synonymy or pertonymy indicate ELABORATION-CONTINUATION, or ELABORATION-RESTATEMENT. Figs. 2 and 3 show how holonymy (*Deutschland – Süddeutschland, Norddeutschland*) induces ELABORATION-DERIVATION, and pertonymy (*Automatisierung – automatisiert*) ELABORATION-DRIFT.

The discourse parser must be able to perform a lookup in the OWL version of GermaNet.[2] To establish a Prolog interface between the OWL version of GermaNet and the discourse parser, we convert the OWL code into a Prolog fact base using the **Thea** ([11]) OWL Library for Prolog, which in turn uses the SWI-Prolog's Semantic Web library[3]. We have implemented Prolog predicates such as `transitive_isHyponymOf_LU(`$LU_1$`,` $LU_2$`,` $Degree$`)`, so that, using the `findall/3`-construction, queries for the sets of direct/transitive hyponyms/hyperonyms can be straightforwardly formulated in Prolog. Corresponding predicates for the remaining lexical-semantic relations allow for further queries. The implemented series of predicates considers the levels of synsets, of lexical units, of the GermaNet orthographic representations, and of the <lemma> tag of the morphological and syntactic tagger that we employ in our parser.


**Fig. 2.** Holonymy as a cue for ELABORATION-DERIVATION

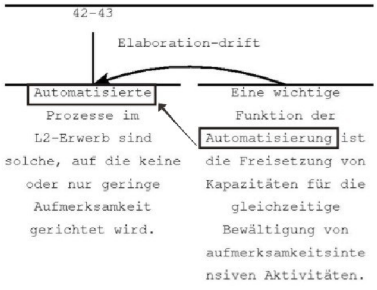


**Fig. 3.** Pertonymy as a cue for ELABORATION-DRIFT

[2] A complete conversion of GermaNet into OWL DL is still pending, so far a sample of GermaNet consisting of 37 synsets and 63 lexical units has been converted [10].

[3] `http://www.swi-prolog.org/`

## 4 Generalised utilisation of OWL ontologies in the GAP

We consider the process of DP as an iterative application of a more general parser architecture which accepts different annotation layers as input data and produces a new annotation layer as its output, see Fig. 4. In each of the consecutive instantiations of the so-called *Generalised Annotation Parser* (GAP), a different set of resources is employed to conrol it.

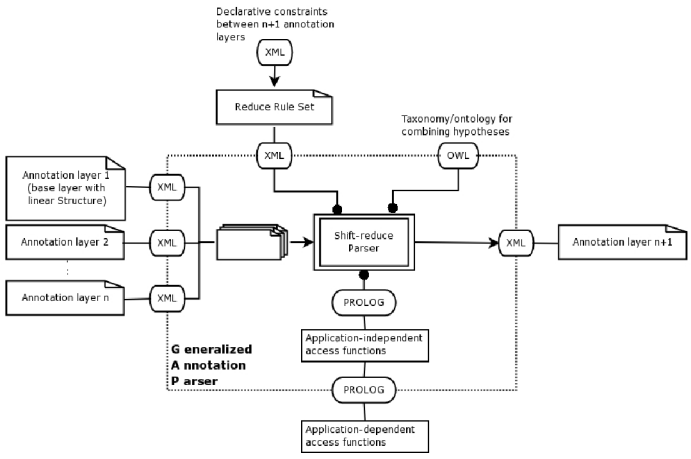


**Fig. 4.** Generalised Annotation Parser GAP

The core of the GAP is a shift-reduce parser, implemented in Prolog. It gets the primary textual data and their $n$ XML annotation layers as its input, which are first converted to a Prolog fact base. The behaviour of the parser is controlled by a set of application-dependent reduce rules formulated in XML. The conditions of their application are expressed as declarative constraints between the $n+1$ annotation layers. The conditions for several subcases of ELABORATION relations expressed in Sect. 3, for example, are formulated as XML reduce rules.

The XML reduce rules set is converted to Prolog rules by the GAP, so that they can directly be used by the shift-reduce parser. The constraints that are part of the reduce rules make use of access predicates which express connections between different annotation layers. The set of access predicates can be divided into application-independent ones, such as `identity(`$layer_i$`:`$element_x$`,` $layer_j$`:`$element_y$`)` or `text-inclusion(`$textvalue$`,` $layer_i$`:`$element_x$`)`, and application-dependent ones which can refer to the schema information of annotation layers.

In many parsing applications it can happen that more than one reduce rule is applicable in a reduce step. Such situations depend on the one hand on the reduce rule set and on the other hand on the structure of the input annotation layers. They lead to competing hypotheses about the combination of segments and therefore to a *set* of possible output annotation hierarchies.

A set of competing but in some way matching hypotheses can be combined by combination rules. In the GAP, such combination rules are derived from the OWL *subclassOf* property that holds between classes of an application-dependent OWL DL ontology.

In the case of DP, whenever two or more competing hypotheses about relation instances have been emitted in the parsing process, the parser consults the RRSet ontology (Sect. 2) and checks whether the $n$ relation names of the competing hypotheses have one or more lowest common superclasses within a certain range, for example within the so-called *reduced relation set*. For each lowest common superclass found, the hypotheses are merged into one, and the superclass is taken as the relation label of the new hypothesis, representing an underspecified relation instance. Like the OWL ontology of GermaNet, the RRSet ontology is converted to Prolog and consulted by the parser using **Thea** ([11]). The final paper will contain the description of an example of competing hypotheses and how it is processed in the GAP.

## References


1. Mann, W.C., Taboada, M.: RST – Rhetorical Structure Theory. W3C page (2005) `http://www.sfu.ca/rst`.
2. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge, MA (2000)
3. Smith, M.K., Welty, C., McGuiness, D.L., (eds.): OWL Web Ontology Language guide. Technical report, W3C recommendation (2004) `http://www.w3.org/TR/2004/REC-owl-guide-20040210`.
4. Hovy, E., Maier, E.: Parsimonious or profligate: How many and which discourse structure relations? Unpublished paper, `http://www.isi.edu/natural-language/people/hovy/publications.html` (1995)
5. Carlson, L., Marcu, D.: Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA (2001) ISI-TR-545.
6. Lüngen, H., Lobin, H., Bärenfänger, M., Hilbert, M., Puskàs, C.: Text parsing of a complex genre. In: Proceedings of the Conference on Electronic Publishing (ELPUB), Bansko, Bulgaria (2006) 247–256
7. Goecke, D., Lüngen, H., Sasaki, F., Witt, A., Farrar, S.: GOLD and discourse: Domain- and community-specific extensions. In: Proceedings of the 2005 E-MELD-Workshop, Boston, MA. (2005)
8. Noy, N., Rector, A., (eds.): Defining n-ary relations on the semantic web. Technical report, W3C Working Group Note (2006) `http://www.w3.org/TR/swbp-n-aryRelations`.
9. Kunze, C.: Lexikalisch-semantische Wortnetze. In Carstensen, K.U.e.a., ed.: Computerlinguistik und Sprachtechnologie: eine Einführung. Spektrum Verlag, Heidelberg (2001) 386–393
10. Kunze, C., Lemnitzer, L., Lüngen, H., Storrer, A.: Modellierung und Integration von Wortnetzen und Domänenontologien in OWL am Beispiel von GermaNet und TermNet. In: Proceedings der KONVENS 2006, Konstanz (2006) To appear.
11. Vassiliadis, V.: Thea. A web ontology language - OWL library for [SWI] Prolog. Web-published manual, `http://www.semanticweb.gr/TheaOWLLib/index.htm`, visited 15.7.2006 (2006)