Maja Bärenfänger, Mirco Hilbert, Henning Lobin, Harald Lüngen

OWL ontologies as a resource for discourse parsing

1 Introduction

In the project SemDok (*Generic document structures in linearly organised texts*) funded by the German Research Foundation DFG, a discourse parser for a complex type (scientific articles by example), is being developed. Discourse parsing (henceforth DP) according to the Rhetorical Structure Theory (RST) (Mann and Taboada, 2005; Marcu, 2000) deals with automatically assigning a text a tree structure in which discourse segments and rhetorical relations between them are marked, such as Concession. 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 lexico-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 (Smith et al., 2004). 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 Taxonomies of rhetorical relations

Already in the original conception of Rhetorical Structure Theory by Mann and Thompson (1988), (see also Mann and Taboada, 2005), 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. Mann and Taboada, 2005). Lower-level subgroups such as *Evidence-and-Justify* were introduced as well. The complete hierarchy is shown in Figure 1.

Hovy and Maier (1995) 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

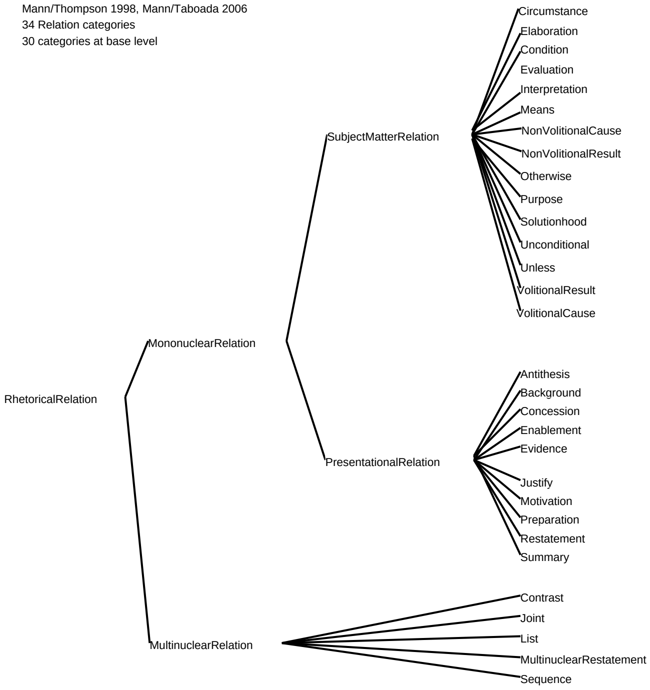


Figure 1: Hierarchy of rhetorical relations according to Mann and Thompson (1988)

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. This hierarchy can be seen in Figure 2

In the present project, we produced corpus annotations using the original RST relation set proposed in Mann and Taboada (2005), and by an examination of these annotations and an inspection of alternative relation sets proposed in the literature (notably Carlson and Marcu (2001) and Hovy and Maier (1995)), we designed a relation hierarchy suitable for annotating the rhetorical structure of scientific journal articles in our explorative reading scenario (Lüngen et al., 2006). 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 an earlier publication (Goecke et al., 2005), we finally refrained

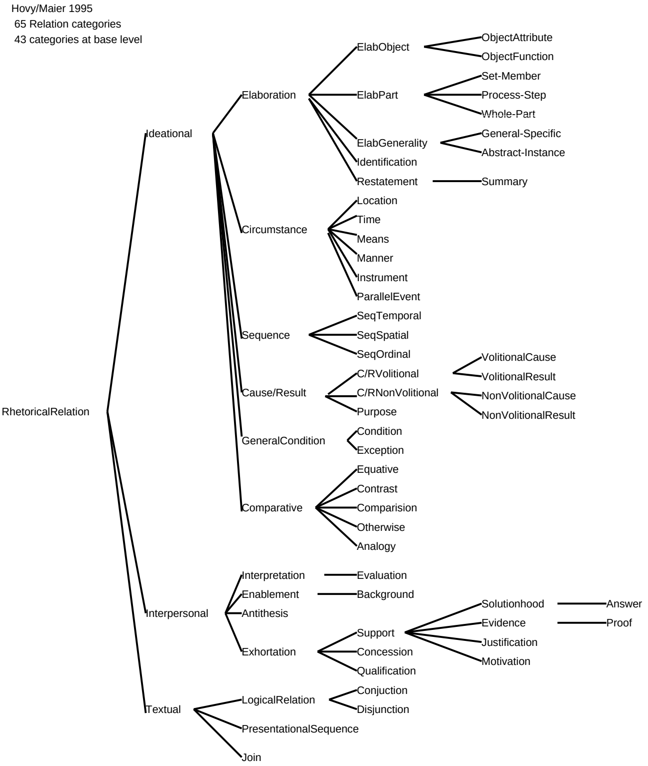


Figure 2: Hierarchy of discourse relations according to Hovy and Maier (1995)

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 (Smith et al., 2004).[1] Thus we modelled the rhetorical relations as OWL classes, which is not so devious if one considers that it is sometimes recommended as good practice to introduce a "relation class" for the encoding of an n-ary relation in OWL (cf. Noy et al., 2006). Subrelation-hood is then marked by the <`rdfs:subclassOf`> construct. The use of <`rdfs:subclassOf`> also enabled us to include further features

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

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 a subclass of InterpersonalRelation as well as of MononuclearRelation. (For reasons of decipherability, the links from MononuclearRelation and MultinuclearRelation are not shown in Figure 3, though.) We introduced further sub- or superrelations, when it was expedient according to our corpus analyses and with respect to our scenario (cf. Lüngen et al., 2006). The resulting hierarchy is shown in Figure 3. 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 of 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. Carlson and Marcu, 2001; Hovy and Maier, 1995; Knott et al., 2001). This led us to distinguish the different subtypes of Elaboration relations which can be seen in the taxonomy of rhetorical relations in Figure 4.

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>), 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).

However, the majority of Elaboration subtypes is not indicated by discourse markers or structural cues, but may be established by the presence of lexical-semantic relations between the central discourse entities of two discourse segments. Elaboration-derivation is signalled by conceptual relations like hyperonymy/ hyponymy, holonymy or meronymy, and lexical relations like synonymy or pertainymy indicate Elaboration-continuation, or Elaboration-restatement. Figures 5 and 6 show how holonymy (*Deutschland – Süddeutschland, Norddeutschland*) induces Elaboration-derivation, and pertainymy (*Automatisierung – automatisiert*) Elaboration-drift.

For the automatic identification of these subtypes there are two options: 1. Lexical-semantic relations may be identified in the discourse parser by performing a lookup

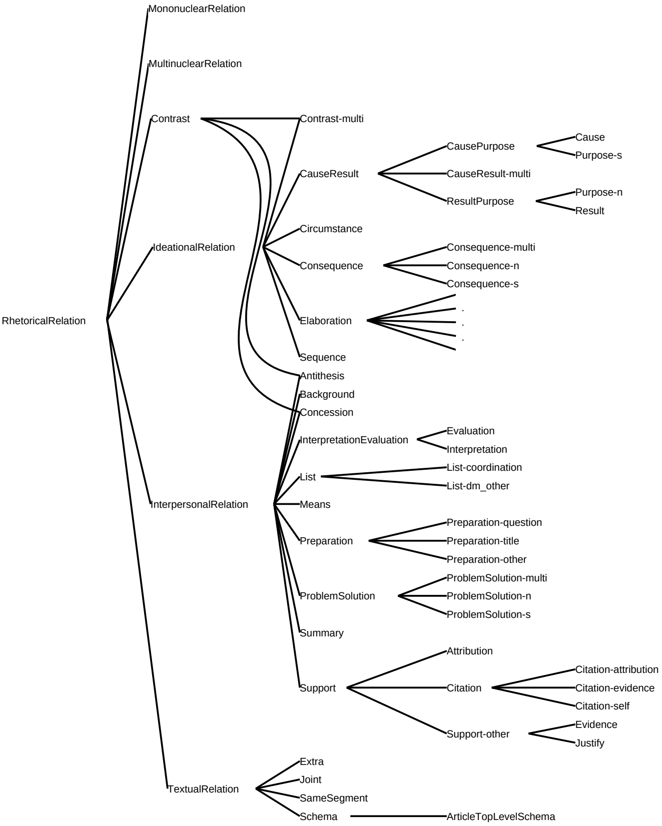


Figure 3: SemDok RRSET ontology (save the subclasses of ELABORATION)

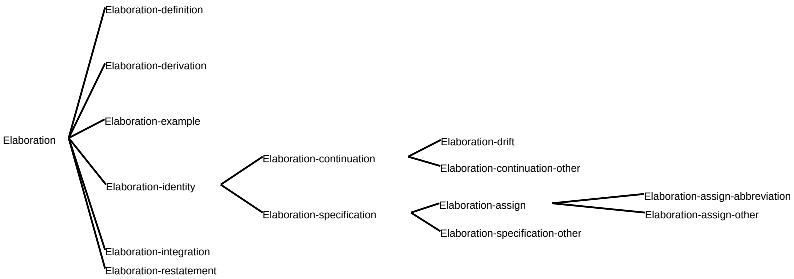
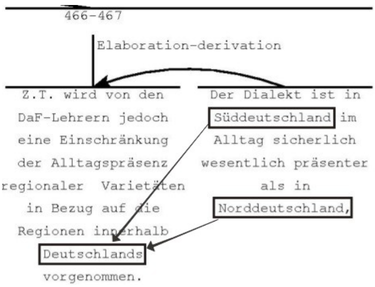


Figure 4: SemDok hierarchy of ELABORATION relations)

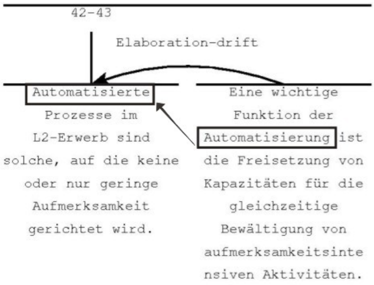
in an OWL version of the lexico-semantic net GermaNet (Kunze et al., 2007). In this approach, GermaNet is directly consulted from the parser. 2. Lexical-semantic relations may be calculated in auxiliary components and be made available to the parser in the form of additional annotation layers of the input text. As auxiliary components, we envisage a lexical chainer and/or an anaphora resolution component as developed in out partner projects HyTex (Holler et al., 2004), and Sekimo (Goecke et al., this volume). As the coverage of our corpus by GermaNet 5.0 seems not high enough for a direct approach – 30.74% of all noun tokens and 59.17% of all noun types in our corpus are not contained in GermaNet – we will first focus on the second option.



Z.T. wird von den DaF-Lehrern jedoch eine Einschränkung der Alltagspräsenz regionaler Varietäten in Bezug auf die Regionen innerhalb Deutschlands vorgenommen.

Der Dialekt ist in Süddeutschland im Alltag sicherlich wesentlich präsenter als in Norddeutschland,

Figure 5: Holonymy as a cue for ELABORATION-DERIVATION



Automatisierte Prozesse im L2-Erwerb sind solche, auf die keine oder nur geringe Aufmerksamkeit gerichtet wird.

Eine wichtige Funktion der Automatisierung ist die Freisetzung von Kapazitäten für die gleichzeitige Bewältigung von aufmerksamkeitsinte nsiven Aktivitäten.

Figure 6: Pertainymy as a cue for ELABORATION-DRIFT

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 Figure 7. In each of the consecutive instantiations of the so-called *Generalised Annotation Parser* (GAP), a different set of resources is employed to control it.

The core of the GAP is a bottom-up passive chart parser, implemented in Prolog. It takes 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, which, for the most part, are derived from a discourse marker lexicon. 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 reduce rules. The reduce rules set is converted to Prolog, so that they can directly be used by the chart parser. The constraints that are part of the reduce rules make

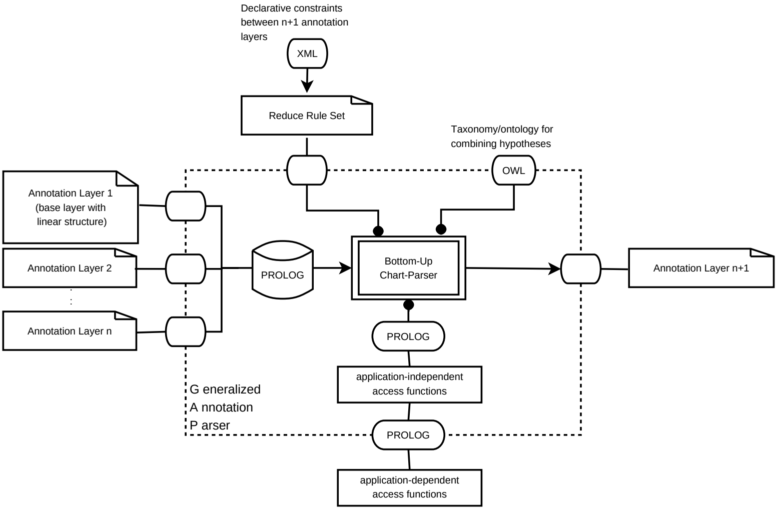

Declarative constraints between n+1 annotation layers
XML
Reduce Rule Set
Taxonomy/ontology for combining hypotheses
OWL
Annotation Layer 1 (base layer with linear structure)
Annotation Layer 2
:
Annotation Layer n
PROLOG
Bottom-Up Chart-Parser
Annotation Layer n+1
PROLOG
application-independent access functions
G eneralized
A nnotation
P arser
PROLOG
application-dependent access functions

Figure 7: Generalised Annotation Parser GAP

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.

As in most 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, specifically, when there are ambiguous discourse markers, such as the German conjunction *aber*, which, similar to English "but" can signal CONCESSION or CONSTRAST-MULTI (cf. Figures 8 and 9[2]). If such an ambiguity cannot be resolved e.g. because no further, supporting discourse markers are present, it leads to competing hypotheses about the combination of segments and therefore to a *set* of possible output annotation hierarchies (two in case of the example). This has two types of negative consequences:

1. In a bottom-up parsing approach, which is mostly taken in DP, the number of alternatives that have to be pursued increases, thus reducing parsing efficiency in terms of time and memory.

[2] Text segments in Figures 5, 8, and 9 are from Baßler and Spiekermann (2001); text segments in Figure 6 are from Bärenfänger and Beyer (2001).

2. When parsing results are evaluated against reference annotations of discourse structures, the non-matching hypothesis will count as an ordinary recall error, although CONSTRAST-MULTI and CONCESSION are semantically quite similar.

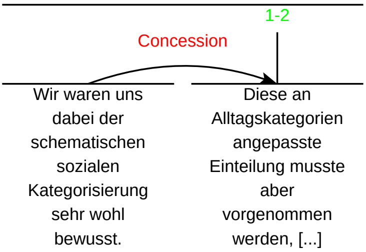

1-2
Concession
Wir waren uns dabei der schematischen sozialen Kategorisierung sehr wohl bewusst.
Diese an Alltagskategorien angepasste Einteilung musste aber vorgenommen werden, [...]

Figure 8: *aber* signalling CONCESSION

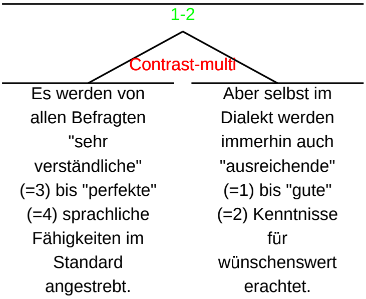

1-2
Contrast-multi
Es werden von allen Befragten "sehr verständliche" (=3) bis "perfekte" (=4) sprachliche Fähigkeiten im Standard angestrebt.
Aber selbst im Dialekt werden immerhin auch "ausreichende" (=1) bis "gute" (=2) Kenntnisse für wünschenswert erachtet.

Figure 9: *aber* signalling CONTRAST-MULTI

Besides introducing local ambiguity packing (Tomita, 1987), the first situation can be remedied by replacing the two hypotheses by one hypothesis with the label of the lowest common superordinate relation according to the RRSET hierarchy, which is CONTRAST in the the example. Such a combination rule can be derived from the OWL *subclassOf* property that holds between classes of an application-dependent OWL DL ontology. 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 check whether the n relation names of the competing hypotheses have one or more lowest common superclasses within a certain range, e.g. within a 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.

In the second situation, in order to differentiate between hard-core recall errors and those caused by semantically similar relations that have been recognised at the same time, an additional evaluation can be conducted where relation labels in the parsing results as well as in the reference annotations are first replaced by labels from a reduced relation set, as e.g. done in Soricut and Marcu (2003). Such a replacement can also be effected by a look-up in the RRSET ontology.

Like the OWL ontology of GermaNet, the RRSet ontology is converted to Prolog and consulted by the parser using the Thea OWL Library for Prolog (Vassiliadis, 2006), which in turn uses the SWI-Prolog's Semantic Web library[3]

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

5 Conclusion

In this article, we sketched the SemDok RRSet relation taxonomy for rhetorical relations in scientific journal articles which was designed based on corpus investigations and previously proposed hierarchies of discourse relations. We described how it was coded in the Web Ontology Language OWL, and how the OWL-based ontology will be consulted as a knowledge base by a discourse parser. As a second example of the utilisation of ontologies in discourse parsing, methods to identify subtypes of the ELABORATION relation using an OWL version of the lexico-semantic net GermaNet were described.

In the GAP, local ambiguity packing is currently employed rather than looking up the RRSET ontology during parsing. However, the RRSET will be used in the evaluation of parsing results as described, and is also used to generate a relations file for manual annotations of discourse structures using O'Donnells RSTTool (O'Donnell, 2000). As for the identification of ELABORATION relations, a comprehensive approach analysing annotations of anaphora and lexical chains is pursued.

References

Bärenfänger, O. and Beyer, S. (2001). Zur Funktion der mündlichen L2-Produktion und zu den damit verbundene kognitiven Prozessesn für den Erwerb der fremdsprachlichen Sprechfertigkeit. *Linguistik Online*, 8. `http://www.linguistik-online.de`.

Baßler, H. and Spiekermann, H. (2001). Dialekt und Standardsprache im DaF-Unterricht. Wie Schüler urteilen - wie Lehrer urteilen. *Linguistik Online*, 9. `http://www. linguistik-online.de`.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA. ISI-TR-545.

Goecke, D., Lüngen, H., Sasaki, F., Witt, A., and Farrar, S. (2005). GOLD and discourse: Domain- and community-specific extensions. In *Proceedings of the 2005 E-MELD-Workshop*, Boston, MA.

Holler, A., Maas, J.-F., and Storrer, A. (2004). Exploiting coreference annotations for text-to-hypertext conversion. In *Proceeding of LREC*, volume II, pages 651–654. Lisboa.

Hovy, E. and Maier, E. (1995). Parsimonious or profligate: How many and which discourse structure relations? Unpublished paper, `http://www.isi.edu/natural-language/people/hovy/publications.html`.

Knott, A., Oberlander, J., O'Donnell, M., and Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In Sanders, T., Schilperoord, J., and Spooren, W., editors, *Text representation: Linguistic and psycholinguistic aspects*, volume 8 of *Human Cognitive Processing*, pages 181–196. Benjamins, Amsterdam.

Kunze, C., Lemnitzer, L., Lüngen, H., and Storre, A. (2007). Repräsentation und verknüpfung allgemeinsprachlicher und terminologischer wortnetze in owl. *Zeitschrift für Sprachwissenschaft*. To appear.

Lüngen, H., Lobin, H., Bärenfänger, M., Hilbert, M., and Puskàs, C. (2006). Text parsing of a complex genre. In *Proceedings of the Conference on Electronic Publishing (ELPUB)*, pages 247–256, Bansko, Bulgaria.

Mann, W. C. and Taboada, M. (2005). RST – Rhetorical Structure Theory. W3C page. `http://www.sfu.ca/rst`.

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

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.

Noy, N., Rector, A., and (eds.) (2006). Defining n-ary relations on the semantic web. Technical report, W3C Working Group Note. `http://www.w3.org/TR/swbp-n-aryRelations`.

O'Donnell, M. (2000). RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pages 253 – 256, Mitzpe Ramon, Israel.

Smith, M. K., Welty, C., McGuiness, D. L., and (eds.) (2004). OWL Web Ontology Language guide. Technical report, W3C recommendation. `http://www.w3.org/TR/2004/REC-owl-guide-20040210`.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edmonton, Canada.

Tomita, M. (1987). An efficient augmented-context-free parsing algorithm. *Computational Linguistics*, 13(1-2):31–46.

Vassiliadis, V. (2006). Thea. A web ontology language - OWL library for [SWI] Prolog. Web-published manual, `http://www.semanticweb.gr/TheaOWLLib/index.htm`, visited 15.7.2006.