Published in: Proceedings of the E-MELD Workshop on Morphosyntactic Annotation and Terminology: Linguistic Ontologies and Data Categories for Language Resources. Electronic Resource - Boston, 2005.

GOLD and Discourse: Domain– and Community–Specific Extensions

Daniela Goecke (Bielefeld University)
Harald Lüngen (Justus–Liebig–Universität Gießen)
Felix Sasaki (World Wide Web Consortium)
Andreas Witt (Bielefeld University)
Scott Farrari (University of Bremen)

1 Introduction

1.1 Aims of the Current Project

Corpora annotated for discourse–related phenomena have become an important source for the empirical study of whole texts and a source of training data for the automated parsing of whole texts. By discourse–related phenomena, we refer to various textual relations (e.g., anaphoric and rhetorical relations) that hold among textual units (e.g., whole texts, adjacency pairs, and discourse segments). One of the main problems associated with corpus linguistics—and for the markup of linguistic data in general—has been the lack of interoperability between disparately marked up resources. While it is theoretically possible to standardize the content and structure of markup, the richness of the data itself suggests that no one standard scheme would suffice for all data, all of the time. In fact one of the major contributions of the E–MELD project[1] has been to show that annotation elements need not be standardized as such. Rather, by following certain parameters of "best–practice" (Bird & Simons 2003) the stage can be set for wide–spread interoperability among disparate corpora. For example, one of the key strategies of the E–MELD project has been to suggest that all markup elements used in annotating linguistic data (including discourse–related corpora) should be mapped to a semantic resource that defines the meaning of each element. Such a semantic resource for descriptive discourse categories does not exist at present. To rectify this situation, we propose a discourse–specific extension to the General Ontology for Linguistic Description (GOLD), as introduced by Farrar & Langendoen (2003) and explicated in Farrar (forthcoming).

Our paper is structured as follows. First in Section 1.2 we give a brief introduction to GOLD and the problem of achieving interoperability over diverse corpora, in Section 1.3 we motivate our domain–specific extension for discourse-related categories. In Section 2 we describe our ontology of discourse categories and relations and give an the example of its application. Section 3 summarizes the applicability of our extension for the description of endangered language and gives an outlook on future work.

1.2 GOLD: Brief Background

The central aim of GOLD is to specify in a formal language all of the categories necessary for the broad-coverage description of linguistic field data, especially pertaining to endangered languages. Initially, GOLD was narrowly focused on morphosyntax, since data marked up for morphosyntactic categories composes a large percentage of linguistic field data. Thanks to the E–MELD project, GOLD is currently being extended to cover other sub–disciplines. We also note that there is on–going ontological work concerning discourse being carried out by Laure Vieu, Laurent Prévot, and colleagues as a part of the DOLCE project (Masolo et al. 2003) and that extending GOLD for discourse categories should eventually take this work into account.

In terms of content, a sub–ontology for descriptive discourse analysis should make clear what categories are necessary for conceptualizing discourse analysis as a scientific sub–discipline of linguistics. One the one hand the ontology should capture the knowledge of a well–trained linguist. Such knowledge includes, first and foremost, the basic categories (viz. concepts) of the field, including discourse segments, discourse markers, basic rhetorical relations, etc. Key to specifying the basic categories is the rich *axiomatization* of discourse concepts in a formal language. Axiomatization means that concepts are defined using a formal language, such as the Web Ontology Language (known as OWL) (McGuinness and van Harmelen 2004), thus providing more than just textual definitions. The major advantage of having such a resource is that it is then possible to link various markup terms to the axiomatized concepts, and achieve interoperability across corpus resources. For example with rich axiomatization, two different terms can be compared according to what they mean, and not just according to their string content. It may be the case, for instance, that two different marked up corpora use the terms '*discourse unit*' and '*discourse segment*' to mean the same thing, or that two corpora use '*coref*' in two, potentially incompatible ways. When such terms are related to an ontology of discourse concepts, their meaning is made clear and automated processing over different corpora is facilitated.

Such a scheme lends itself to a *two–level* approach to corpus analysis whereby the structure and the content of annotation are treated differently. The need and benefit of having two layers of corpus analysis can also be seen by the fact that two query languages are being developed by the W3C. The first is XQuery (Boag et al. 2005) which is concerned with the analysis of corpus syntax, i.e. document structures, for example, the names, contents and structure of tags, etc. On the other hand, SPARQL (Prud'hommeaux and Seaborne 2005) supplies query mechanisms for RDF / OWL graphs, which can be used to analyze the meaning of corpus units.

1.3 The need of an extension of GOLD with discourse-related categories

As motivation for a discourse–related extension, this paper introduces applications envisaged in the research group "Text–technological modelling of information", funded by the German Research Foundation (DFG). The research group deals with various modeling and processing areas of (mainly textual) marked–up XML data. A central application domain of the research group, and hence to the GOLD effort, is the processing of discourse relations. This applied, for example, to discourse parsing or to the automatic resolution of anaphoric relations by integrating heterogeneous knowledge resources, such as annotated corpora, lexica or ontologies. In these applications, GOLD plays a crucial role with respect to the following two tasks.

(1) GOLD supplies domain–specific knowledge during the processing of scientific articles from the discipline of linguistics which naturally contain a large amount of linguistic terminology. In this task, the role of GOLD can be compared to that of WordNet in the anaphora resolution algorithm proposed by Vieira and Poesio (2001). Their algorithm focuses on Definite Descriptions, i.e. definite noun phrases, which are an important means to signal anaphoric relations to an antecedent, to introduce new entities in the discourse, to mark discourse segments and to trigger the process of bridging inferences (see section 2.1 of this paper). Their algorithm relies on annotated nominal phrases, various heuristics making use of linguistic knowledge, and the WordNet ontology. WordNet encompasses a basic taxonomy of conceptual relations between so–called synsets, which contain a set of lexical units. These relations, once established in GOLD, can be used for the processing of linguistic terminology. Several relations between nouns occurring in a scientific text, or their synsets, respectively, are possible:

- The nouns are in the same synset (i.e., they are synonyms of each other), as in anaphora — anaphoric expressions;
- The nouns are in a hyponymy/hypernymy relation with each other, as in referential relation — linguistic relation;
- There is a direct or indirect meronymy/holonymy (part of/has parts) relation between them, as in noun — noun phrase;
- The nouns are coordinate sisters, i.e. hyponyms of the same hypernym, such as pronoun — common noun, which are hyponyms of anaphoric expression.

Many of these relations cannot be found in WordNet itself, because WordNet does not contain specific knowledge about the domain of linguistics. Hence, the envisaged extension of GOLD can be used for the encoding of these relations.

(2) Cimiano and Handschuh (2003) use an ontology as a backbone for linguistic annotation. They argue that the (theory–specific) differences between corpus annotations can be overcome with their ontology-based approach. As an example, they describe the annotation of anaphoric relations. Cimiano and Handschuh describe a top–down annotation, i.e. from an existing ontology to the creation of corpus annotations. Although this is complementary to GOLD's main application scenario, i.e. building a bridge between existing theory–specific annotations, GOLD can be used in the same way: Specific annotations can be created with the ontology as their common backbone. The differences between the two approaches concern the ontological specification of discourse categories, which are discussed in sec. 2.2 of this paper.

Finally, in addition to providing an infrastructure for the basic categories of discourse, the ontology should provide the mechanisms for extending the basics in directions that may be community or theory specific. Until now automated comparison of corpora that are analyzed according to divergent theories has been impossible, not only due to differences in terminologies, but also due to incompatible conceptualizations. An ontology for discourse would help to bridge the gap between such analyses such that smart search could be performed over disparate corpora. Whereas the basic categories provide a neutral conceptualization of discourse concepts-in so far as this is possible-various community extensions link to this neutral backbone.

2 Ontology of discourse categories and relations

In this section we discuss the domain of discourse analysis and how various discourse concepts and relations can be described in terms of GOLD. Key to our discussion are two types of discourse relations: *anaphoric* and *rhetorical* relations. But in order to discuss these types of relations in an ontological context, we have to be clear as to what kinds of entities are related in the first place. In the following section we clarify the ontological status of these various discourse–related entities and discuss the sub–types of anaphoric and rhetorical relations.

2.1 Anaphoric Relations

Anaphora as a cohesive device occurs when the interpretation of a text element is dependent on the interpretation of another. The anaphoric element is called the *anaphor* and the element on which it depends is called the *antecedent*[2]. The anaphor is often an abbreviated or reformulated reference to its antecedent, e.g. (examples taken from Clark 1977):

(1) I met a man yesterday. He told me a story.
(2) I met a man yesterday. The man told me a story.

For a description of anaphoric relations two different views have to be taken into account: First, anaphoric relations can be said to hold between formal units within the clause, e.g. between a noun phrase and the referring pronoun as in (1). Second, it can be said that anaphoric relations hold between the meanings, i.e., semantic interpretations, of formal units. In the case of NP antecedents, these meanings are referred to as *discourse entities*. Discourse entities—or discourse referents as Karttunen

(1976) calls them—are constants within a discourse model evoked by (mainly definite) NPs and which can be referred to in the subsequent discourse. NPs can either evoke new discourse entities in the discourse model or can refer to ones that are already there" (Webber, 1988). Webber describes the basic features of a discourse entity:

> "(a) it is a constant within the discourse model and [...]
> (b) one can attribute to it [...] properties and relationships with other entries."
> (Webber 1988, p. 113)

According to the first view, a taxonomy can be built on the basis of the syntactic properties of the anaphora (cf. Hirst, 1981; Mitkov, 2002). In our extension of GOLD, these syntactic properties of the anaphora are modeled as the relation ANAPHORICRELATION that holds directly between formal linguistic units (SYNTACTICUNIT), i.e. between antecedent and anaphor (the domain and range of ANAPHORICRELATION is SYNTACTICUNIT). In the current extension we focus mainly on Germanic languages, however some anaphoric relations for non–Indoeuropean languages as Japanese or Kilivila (SIL code KIJ, see Section 2.3 for a further discussion of Kilivila) have also been included. Figure 1 shows the subclasses of ANAPHORICRELATION.

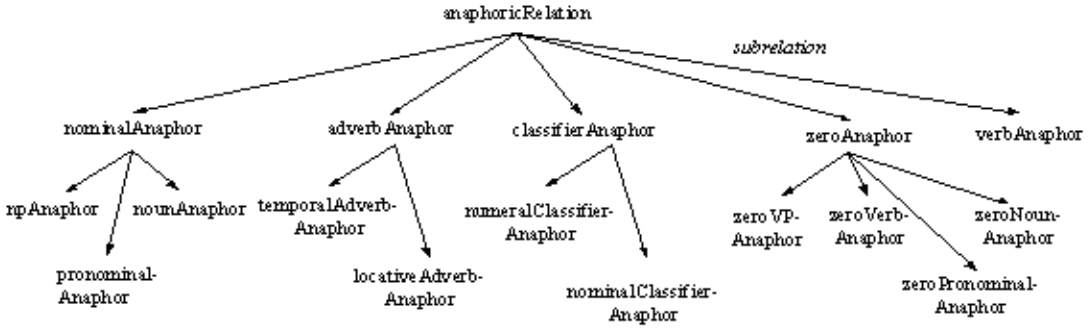

Figure 1: hierarchy of ANAPHORICRELATION

In addition to ANAPHORICRELATION, anaphora can be modelled in terms of relations that hold between the semantic interpretations that are encoded by the linguistic expressions. In the current version of our extension we focus on discourse entities as semantic interpretations[3]. A noun phrase can either evoke a new discourse entity or can refer to an already existing one. Figure 2 exemplifies the correspondence between linguistic expressions and discourse entities.

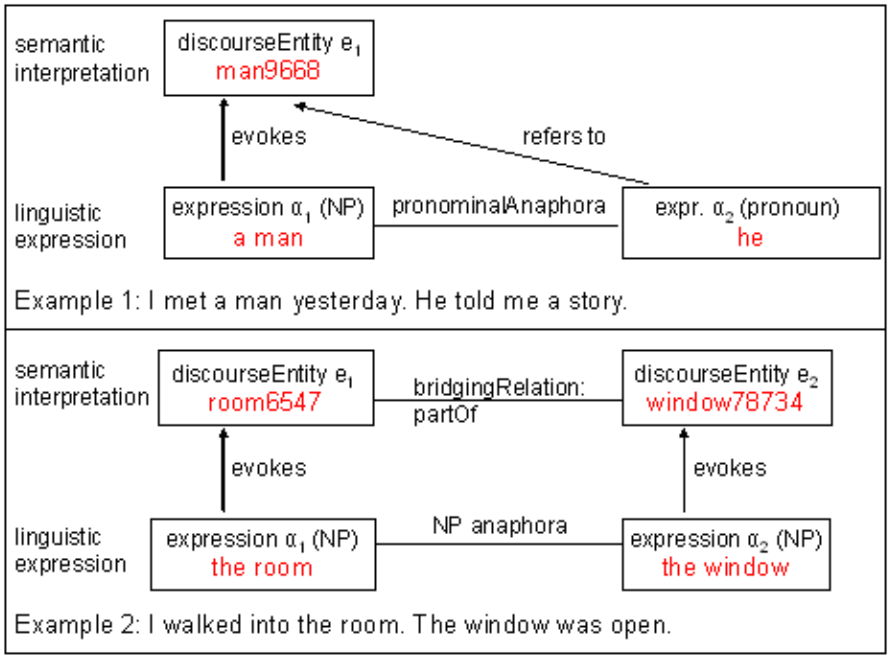

Figure 2: linguistic expressions and semantic interpretation

The first example shows the relation PRONOMINALANAPHOR, in which the pronoun 'he' refers back to its antecedent (the NP 'a man'). The NP 'a man' evokes a new discourse entity in the discourse model. The anaphor refers to the same discourse entity. This type of anaphor is direct anaphor as the antecedent is explicitly mentioned in the context. Furthermore it is of type identity–of–reference as the two linguistic expressions refer to the same discourse entity (i.e. coreference).

The second example is of type NPANAPHOR. Both linguistic expressions evoke (different) discourse entities. This example is indirect anaphor as the antecedent of the second expression α2 is not explicitly mentioned but has to be inferred from the context. On the level of semantic interpretation, this inference can be drawn: Between the two discourse entities e1 and e2 a partOf–relation holds (HOLONYMY in our domain–specific extension). Thus α1 serves — via inference — indirectly as an antecedent for α2.

Based on the relations that hold between the discourse entities, anaphora can be further subdivided into direct anaphora and indirect anaphora.

- direct anaphora: the antecedent is explicitly mentioned in the context (coreference).
- indirect anaphora: the antecedent is not mentioned explicitly in the text but has to be inferred from the context (bridging).

The semantic interpretations have to be taken into account in order to distinguish between direct and indirect anaphora and to trigger the resolution of anaphora involving *definite descriptions* (i.e. NPANAPHORA). Clark (1977) uses the term *bridging*, which he introduces as "the construction of [...] implicatures" and as "an obligatory part of the process of comprehension" (p. 413). *Bridging references* occur when the antecedent of a definite description is not mentioned explicitly in the text but has to be inferred from the context. Clark (1977) gives a taxonomy of bridging references based on relations between the semantic interpretations of the linguistic expressions. Vieria and Teufel (1997) classify bridging descriptions according to the kind of information that is needed to resolve them. According to Webber (1988), direct anaphora (coreference) and indirect anaphora (bridging) can be uniformly modeled by saying that the discourse entity e encoded by an anaphoric expression a is either equal to or associated via bridging references with an existing discourse referent. Thus, two subtasks can be identified for the resolution of indirect anaphora: to find the antecedent for the anaphoric expression and to identify the relation between the discourse entities denoted by the anaphoric expression and its antecedent. We introduce BRIDGINGRELATION as a sub concept of DISCOURSERELATION to cover these relations. Domain and range for BRIDGINGRELATION is DISCOURSEENTITY. The taxonomy for DISCOURSERELATION is given in Figure 3; see Section 2.2. for a description of RHETORICALRELATION.

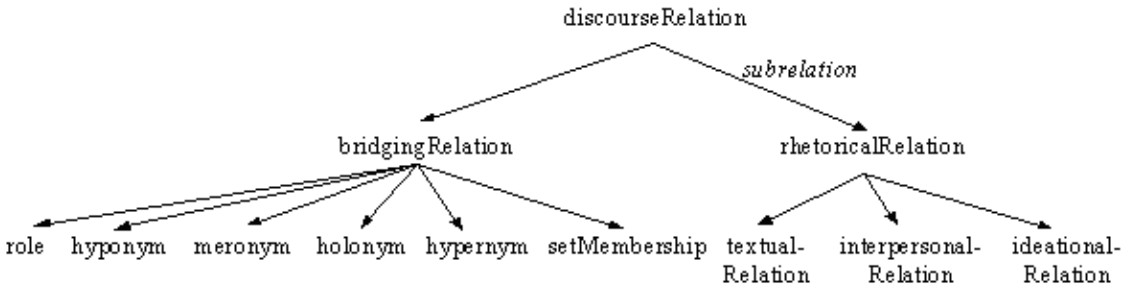

Figure 3: hierarchy of BRIDGINGRELATION and RHETORICALRELATION

The set of ANAPHORICRELATION that has been included in the current version of our extension of GOLD — although not complete yet — provides a basis to classify anaphoric relations according to the morpho–syntactic properties of the anaphoric expression. These relations do not only include categories for romance or germanic languages but form a starting point for other, typologically different languages, e.g. Japanese or Kilivila (see Section 2.3). Regarding DISCOURSERELATION, the current version of our extension allows the description of BRIDGINGRELATIONs between discourse entities. However, there are other subsets of DISCOURSERELATION that fall into the class of direct anaphora but that do not hold between discourse entities denoted by NPs but that imply reference discourse segments which come about through other types of linguistic expressions, e.g. clauses, sentences or even larger text units. According to Eckert and Strube (2001), in spoken language less than half of the occurrences of pronominal anaphora have an NP antecedent. On the level of ANAPHORICRELATION, these types of anaphora can already be described but on the level of DISCOURSERELATION, additional relations have to be defined[4].

2.2 Rhetorical Relations

Rhetorical relations such as introduced in Rhetorical Structure Theory (RST, Mann and Thompson 1988) are often equated with discourse relations proper. They are relations between adjacent portions of text which are rhetorical and/or semantic in nature and, like anaphoric relations, they bring about textual coherence (cf. Hovy and Maier, 1995). Thus, for the two sentences

(3) S1[Peter cannot come to the meeting today]. S2[He is sick.]

to form a coherent text, most readers would infer that the discourse relation *CAUSE(S2,S1)* holds between S1 and S2, although it is not explicitly indicated by a discourse connective such as because or as. From recent publications on discourse theory or discourse annotation, two different views about the domain and range of discourse relations can be made out. The first view is that discourse relations hold directly between portions of text that are *linguistic expressions* such as clauses, sentences, or larger text units. The second view is that discourse relations rather hold between the *semantic interpretations* of these linguistic expressions, such as propositional contents, eventualities, or situations. The first view is prevalent in Mann/Thompson 1988, where it is stated that rhetorical relations hold between "text spans", "text parts", or "clauses", even while they are defined in terms of the situations that are "presented" by the text spans (cf. Mann and Taboada 2005). Also the relation annotation approach by Carlson and Marcu (2001) seems to adhere to such a view as they define elementary discourse units (those units that discourse relations hold between) comprehensively in terms of types of syntactic constituents. The second view is e.g. taken by Webber et al. (2003), where a discourse relation is represented as a logical predicate (that has propositions as its arguments), the source of which may be a discourse connective (a sentence conjunction or adverbial). In many contexts though, the question whether discourse relations are relations between forms or meanings is not considered significant, or left open so that one can find formulations like "adjacency or conjunction [..] imply discourse relations between (*the interpretation of*) adjacent or conjoined discourse units" (Webber et al., 2003, our emphasis). When creating an ontology of such relations, however, this

distinction becomes crucial; that is, the ontological nature of such relations should be made as explicit as possible as to be able to construct statements about them in a formal language. Thus while it is convenient to talk about discourse relations as though they were form relations, in many approaches it is implicitly understood or openly stated that they are relations between the meanings (semantic interpretations) of linguistic forms. Therefore, we take the position that rhetorical relations that corresponds to the latter approach, namely, that rhetorical discourse relations hold between meanings. The approach is illustrated in Figure 4.

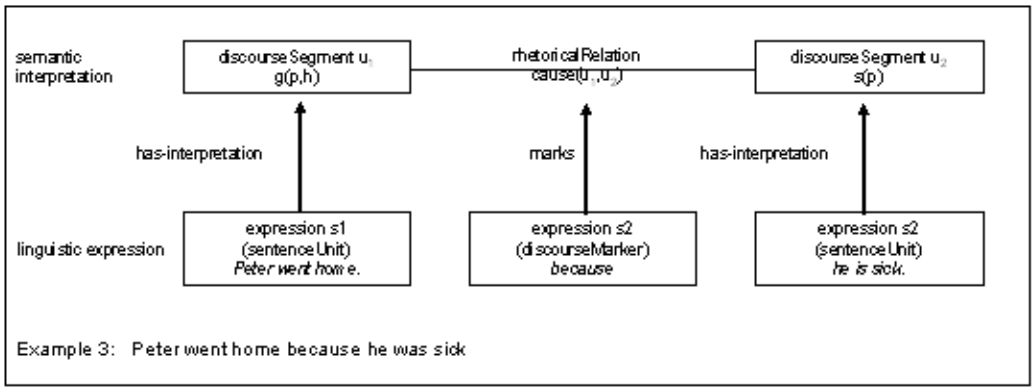

Figure 4: Domain and range of rhetorical relations

To sum up, we propose rhetorical discourse relations to hold between subtypes of discourse semantic units which are represented by the class DISCOURSESEGMENT in our extension of GOLD. Discourse segments, unlike discourse entities, are propositional and are introduced by linguistic expressions that typically encode propositions, such as clauses, sentences, or larger text units—in certain cases also NPs (e.g. when the head noun is a nominalized verb). We introduce DISCOURSEUNIT as a cover concept for DISCOURSESEGMENT and DISCOURSEENTITY, which both are types of semantic concepts, i.e. discourse–relevant roles of semantic interpretations of certain form units. Discourse segments come in various types, for example, those that do not contain other discourse segments as constituents. This corresponds roughly to (the interpretation of) an "elementary discourse unit" (EDU) as in Carlson and Marcu (2001), to a "basic discourse unit" (BDU) as defined in Polanyi et al. (2004), to a "discourse segment" after Alonso Alemany et al. (2004), or to a "basic segment" in Hovy and Maier (1995). We capture this notion in GOLD as the class ELEMENTARYDISCOURSESEGMENT which is a direct of subclass DISCOURSESEGMENT. The other type of discourse segment is composed of other segments and is, thus, not elementary. We capture this notion in GOLD as the class COMPLEXDISCOURSESEGMENT and declare instances of this class to be disjoint from instances of ELEMENTARYDISCOURSESEGMENT (cf. Figure 5).

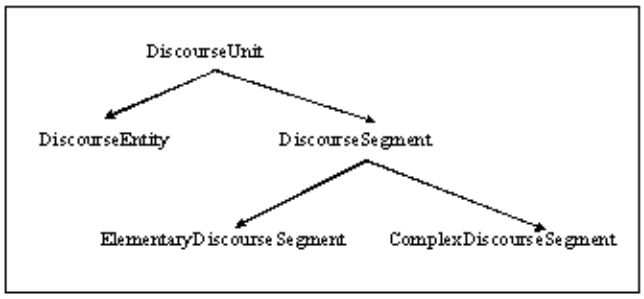

Figure 5: Taxonomy of concepts under DISCOURSEUNIT

What then is the place of "discourse markers" (Alonso Alemany et al., 2004), or "operator segments" (Polanyi et al., 2004), or "discourse connectives" (Korelsky and Kittredge, 1993; Webber et al., 2003), or "cue words and phrases" (Hovy and Maier, 1995) in the ontology? These may be conjunctions like *although*, *because*, or *but*, or adverbs like then or otherwise, but also phrases like *in order to*, or sentences such as *The following summarizes the arguments presented above*. They are said to "indicate", or "mark", or "convey" rhetorical relations. Following the analyses given in Webber et al. (2003), such discourse connectives are linguistic expressions (i.e. signs) that introduce logical predicates representing discourse relations on the discourse semantic level, cf. Fig. 4.[5,6], Thus, our class DISCOURSEMARKER is seen as a role that certain types of sign can assume when considered from a discourse analysis perspective.

Having placed discourse (rhetorical) relations in the ontology by explicating their relationships with SEMANTICUNITS, SIGNS, DISCOURSEENTITIES, and DISCOURSEMARKERS, let us now turn to the question of classifying rhetorical relations into further subrelations. The classification that Mann and Thompson chiefly suggest, is into *subject matter relations* and *presentational relations*, where subject matter relations are those that "express part of the subject matter of the text", e.g. two discourse segments might be causally related in the subject matter, while presentational relations, such as JUSTIFICATION, "only facilitate the presentation process itself" (Mann and Thompson, 1988, p.256). This subdivision is also mirrored in Moore and Pollack (1992), who suggest that texts should actually be analysed simultaneously on two separate discourse levels called the *informational* and the *intentional* level of discourse. They identify subject–matter relations with relations on the informational level and presentational relations with relations on the intentional level. These two categories can also be made out in the three–way top–level partition of rhetorical relations in Hovy and Maier (1995). Following Halliday's (1985) subcategorization of linguistic

phenomena, they subdivide rhetorical relations into ideational (i.e. semantic/subject–matter/informational), interpersonal (i.e. argumentative/presentational/intentional), and textual relations. Ideational relations are relations "relations that express some experience of the world about us and within our imagination" (Hovy and Maier, 1995). Typical ideational relations found in the literature are ELABORATION, CIRCUMSTANCE, or SEQUENCE. Interpersonal relations are those that make reference to and influence the state of mind of the reader in some way. JUSTIFICATION, MOTIVATION, and CONCESSION are typical interpersonal relations. The third category of textual relations was not previously envisaged by Mann and Thompson (1988) or Moore and Pollack (1992). They are "holding between adjacent segments of text that are not meant to be directly related ideationally or interpersonally, but whose relationship exists solely due to the juxtaposition imposed by the nature of the presentation medium" (Hovy and Maier, 1995). An example of such a relation is PRESENTATIONALSEQUENCE. Consequently, we propose to partition rhetorical relations immediately into ideational, interpersonal, and textual relations, following the analyses and definitions in Hovy and Maier (1995). The paper by Hovy and Maier (1995), which is much-cited but so far unpublished (other than via the web), actually includes a well–crafted taxonomy of rhetorical relations based on a comprehensive review of previous literature and projects, with special emphasis on a text generation point of view. We would consider this as an example of a COPE (community–specific extension), which can now easily integrated in the present ontology if desired. Thus, we think the present ontology of discourse relations down to the level of ideational, interpersonal, and textual relations is what linguists working with discourse would agree on, and project–specific relation sets or hierarchies can now be 'inserted' below this level. The discussed top–level of the hierarchy can be seen in Figure 3.

2.3 Kilivila-Data

This section examines the application of the proposed discourse categories to the markup of Kilivila data. We have chosen Kilivila primarily because it illustrates the complexity of anaphoric and referential relations that can be encountered in lesser studied (in fact endangered) languages.

Typologically different languages use different linguistic means to express anaphoric relations. Japanese, for example, often uses zero-pronouns or numeral classifiers to express coreference, i.e. reference identity between anaphor and antecedent. In Kilivila, classificatory particles are one possibility to express reference identity, whereas Germanic languages most often use pronouns (cf. Sasaki et al., 2002). In Kilivila, anaphoric and referential relations can be expressed in a number of ways (cf. Senft, 1986). Pronouns can be used to express anaphoric relations. However, this is quite uncommon as the use of pronouns has certain discourse functions, e.g. emphasis. Much more often the nominal referent is omitted after having been introduced (zero anaphora). A technique of nominal classification is employed: Several word classes (including demonstrative pronouns, numerals and some adjectives) have to be marked with respect to the class of the noun they refer to. This is done by attaching or inserting 'classificatory particles' (CPs, cf. Senft, 1996). The corresponding CP is used to refer to the omitted referent, even beyond sentence boundaries, and sometimes several utterances after the introduction of this referent, thus securing coherence in discourse. The following example, taken from Sasaki et al. (2002), exemplifies the function of the CPs in establishing anaphoric relations:

ke$_i$-	*ta*	***kai***$_i$	*ku-*	*kau*	…	***ke***$_i$-	*bwabwau*
CP.wooden-one		stick	2.-	take		CP.wooden-	blue
'take one stick…'						'the blue (stick)…'	

Example: Classificatory Particles in Kilivila

In the first part of the discourse, the noun kai 'stick' is explicitly mentioned. The discourse entity that is evoked by the noun phrase is taken up later in what is the second part of the example, but is now only referred to by a nominal phrase consisting of an adjective containing the CP *ke* 'wooden'. The noun is not mentioned again. The CP also occurs in the first part of the example, attached to a numeral modifying the noun *kai* 'stick'. The occurrence of this CP in both nominal phrases bridges the gap between them and creates a referential relation. Translated to our extension of the GOLD ontology, the above example is of type NOMINALCLASSIFIERANAPHOR. In addition, constraints on antecedent–anaphor relations have to be formulated to ensure identity of the classifier particle type ('wooden' in the above example). Similar restrictions have to be formulated for other languages, too, e.g. to ensure number or gender agreement for PRONOMINALANAPHOR (cf. Mitkov 2002).

3 Summary and Outlook

The purpose of this paper was two-fold. First, it was shown that a domain–specific extension, can be integrated into the GOLD framework. This allows for making use of the core GOLD ontology and to combine it with additional theory–driven linguistic knowledge for other domains of application. In this paper, linguistic knowledge especially used in the field of computational linguistics and language technology has been described as a domain–specific extension. Second, as has been shown for coreference in Kilivila, the description of lesser known languages can benefit from a domain–specific extension, too. This offers new perspectives for a broader description of endangered languages.

To these ends, we have presented an extension of the GOLD ontology for discourse–related categories. The focus was on discourse coherence relations, which come in two types, that is, anaphoric relations and rhetorical relations. The domain and range of anaphoric relations are nominal linguistic expressions (syntactic units) while they are brought about either by co–reference or by bridging relations between discourse entities on the semantic level. Rhetorical relations hold between discourse segments which are the semantic interpretations of text units. They are classified into ideational, interpersonal, and textual

relations. Finally, discourse markers are introduced as a subclass of sign that denote rhetorical relations on the discourse semantic level. We intend to develop a project specific set of rhetorical relations that is tuned to the text analysis of scientific articles in an e–learning scenario. This set will be hierarchically structured and will form a COPE to the present ontology.

Further work is needed to extend the class ANAPHORICRELATION in order to add at least the major subclasses for a variety of languages. Additionally, certain restrictions for antecedent–anaphor relation have to be included, e.g. gender and number agreement for NOMINALANAPHOR or an identity restriction for classificatory particles. We plan to model these restrictions as ANAPHORICRESTRICTION.

1. E-MELD is a NSF–funded project that supported GOLD. See the website, http://emeld.org

2. Anaphor is when some text element refers back to a previous one, while cataphor presupposes a following item. However, as Halliday and Hasan (1976) point out, this "distinction only arises if there is an explicitly presupposing item present, whose referent clearly either precedes or follows. If the cohesion is lexical, with the same lexical item occurring twice over, then obviously the second occurrence must take its interpretation from the first; the first can never be said to point forward to the second." (p.17). Work of the DFG research group focuses on definite descriptions, thus cataphor is not incorporated in our COPE.

3. In English, discourse entities are mostly introduced by definite NPs. Karttunen (1976) also gives examples of indefinite NPs introducing discourse entities.

4. The following example taken from Asher (1993) is of type PRONOMINALANAPHOR. The antecedent, however, does not denote a discourse entity of individual type but "a discourse referent p1of propositional type" (p.241): John believes that [Mary is a genius]i. Fred is certain of it i.

5. The gist of Webber et al. (2003) is actually that there are some discourse connectives (namely adverbial ones) who find one of their arguments through anaphora (not through adjacency). For the time being, this view is not integrated in our ontology.

6. Note that while discourse markers convey discourse relations, discourse relations need not necessarily be conveyed by discourse markers.

References

Alonso Alemany, Laura and Ezequiel Andújar Hinojosa and Robert Sola Salvatierra (2004). A framework for feature–based description of low–level discourse. *In Proceedings of the ACL Workshop on Discourse Annotation*, 1–8, Barcelona, Spain.

Asher, Nicholas (1993). *Reference to Abstract Objects in Discourse. (Studies in Linguistics and Philiosophy, vol. 50)*. Dordrect London: Kluwer Academic Publishers.

Bird, Steven, and Simons, Gary (2003). *Seven Dimensions of Portability for Language*. Language 79.3.557–82. http://www.language–archives.org/documents/portability.pdf

Carlson, Lynn and Daniel Marcu (2001). *Discourse Tagging Reference Manual*. ISI Rech Report ISI–TR–545.

Cimiano, Philipp and Siegfried Handschuh (2003). Ontology–based linguistic annotation. *In Proceedings of the ACL Workshop on Linguistic Annotation*, Sapporo, Japan.

Clark, H. (1977). *Bridging*. In: P.N.Johnson–Laird & P.C.Wason (eds.) Thinking: Readings in Cognitive Science, 411–420, Cambridge University Press, Cambridge.

Eckert, M. and Strube, M. (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, vol. 17, no. 1, pp. 51–89.

Farrar, Scott and D. Terence Langendoen (2003) A linguistic ontology for the SemanticWeb. *GLOT International* 7(3), 97 100.

Farrar, Scott (forthcoming) Using Ontolinguistics for language description . In: A. Schalley and D. Zaeferer (eds.): Ontolinguistics: *How Ontological Status Shapes the Linguistic Coding of Concepts*. Berlin: Mouton de Gruyter.

Halliday, M.A.K. and Hasan, R. (1976). *Cohesion in English*. Longman English Language Series 9. London: Longman.

Halliday, M.A.K. (1985). *An Introduction into Functional Grammar*. Baltimore, Edward Arnold Press.

Hirst, Graeme (1981). *Anaphora in Natural Language Understanding: A Survey*. Berlin, Heidelberg: Springer.

Hovy, Eduard und Elisabeth Maier (1995). *Parsimonious or profligate: How many and which discourse structure relations?* http://www.isi.edu/natural-language/people/hovy/publications.html

Karttunen, Lauri (1976). Discourse referents. *Syntax and Semantics: Notes from the Linguistic Underground*, 7:363–385.

Korelsky, Tanya and Richard Kittredge (1993). Towards stratification of RST. In: O. Rambow (ed.), *Proceedings of the ACL Workshop on Intentionality and Structure in Discourse Relations*. Ohio State University, pp. 52–55.

Kruijff–Korbayová, Ivana and Geert–Jan M. Kruijff (2004) Discourse–level annotation for investigating information structure. *In Proceedings of the ACL Workshop on Discourse Annotation*, 41–48, Barcelona, Spain.

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

Mann, Bill and Maite Taboada (2005). *Relation definitions*. Web document, http://www.sfu.ca/rst/01intro/definitions.html, accessed 2005–05–24.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider: 2002, The WonderWeb library of foundational ontologies: preliminary report . WonderWeb Deliverable D17, ISTC–CNR, Padova, Italy.

McGuinness, D. L. and F. van Harmelen (2004) OWL Web Ontology Language: Overview. W3C Recommendation 10 February 2004. Available at http://www.w3.org/TR/owl–features/.

Mitkov, Ruslan (2002). *Anaphora Resolution*. London: Longman.

Moore, Johanna D. and Martha E. Pollack (1992). A problem for RST: The need for multi–level discourse analysis. *Computational Linguistics*, 18(4):538–544.

Polanyi, Livia and Chris Culy and Martin van den Berg and Gian Lorenzo Thione and David Ahn (2004). A rule–based approach to discourse parsing. *In Proceedings of the 5th Workshop in Discourse and Dialogue*, 108–117, Cambridge, MA.

Sasaki, F., C. Wegener, A. Witt, D. Metzing and J. Pönninghaus (2002). Co–reference annotation and resources: a multilingual

corpus of typologically diverse languages. *In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC–2002)*, Las Palmas.
Senft, G. (1986). *Kilivila: the language of the Trobriand islanders*. Berlin: Mouton de Gruyter.
Senft, G. (1996). *Classificatory particles in Kilivila*. New York: Oxford University Press.
Webber Bonnie (1988). Discourse deixis: Reference to discourse segments. *In Proceedings of the 26th Annual Meeting of the Association for Computational Linguistic (ACL—88)*, pages 113–122, State University of New York at Buffalo, June 27–30 1988.
Webber, Bonnie and Matthew Stone and Aravind Joshi and Alistair Knott (2003). Anaphora and Discourse Structure. *Computational Linguistics*, 29(4):545–587.