

On the Resolution of Bridging References within Information Extraction Systems

Philipp Cimiano Lavin

January 2, 2003

Contents

1	Introduction	3
2	The SWISS-PROT Corpus	6
3	Related Work	10
3.1	Bridging	10
3.1.1	Bridging as Coercive Accommodation	10
3.1.2	Bridging as Byproduct of Discourse Interpretation	15
3.1.3	Interpretation as Abduction	18
3.1.4	Constructing Bridges	18
3.1.5	Empirically Based Approaches	19
3.2	Discourse Analysis in Information Extraction	20
3.3	Ontologies for Molecular Biology	22
3.4	Discussion	22
4	The Ontology Driven Approach	26
4.1	The Role of World Knowledge in Bridging Reference Resolution	27
4.2	Reasoning within an Ontology	33
4.3	Integrating the Ontology	36
4.4	Determinacy and Reasoning Complexity	38
5	The Ontology	40
5.1	Pathways: the General Picture	40
5.2	Event or not event?	44
5.2.1	Linguistic Realization of Events	44
5.2.2	A Semantic Classification of Events	45
5.2.3	The Realistic Picture	59
5.3	Design Principles	59
5.3.1	Ontological Representation of Events	59
5.3.2	Taxonomic Relations	65
5.3.3	Conceptual Definitions	68

6	Application to Examples	70
7	Implementation	82
7.1	anaphora_resolution.pl	82
7.2	ontology.pl	85
7.3	bridging.pl	85
7.4	reasoning.pl	86
7.5	drs_calculus.pl	87
8	Evaluation and Results	88
8.1	Training and Testing	88
8.2	Ontology Development	89
8.3	Agreement between Annotators	90
8.4	Results	92
8.5	Discussion of the Results	94
8.5.1	Annotation Errors	94
8.5.2	Underspecification	94
8.5.3	Recall and Precision	95
8.6	Exploiting Lexical Clues in the Resolution Process	96
8.7	The Turing-test	98
8.8	Conclusions	100
9	Conclusion and Outlook	102
A	Ambiguous Antecedents	105
B	SWI Prolog Source Code	111
B.1	anaphora_resolution.pl	111
B.2	bridging.pl	113
B.3	reasoning.pl	113
B.4	gmp.pl	114
C	The Ontology O_{Bio}	116
C.1	The Concepts in C_{Bio}	116
C.2	The Taxonomic Relations T_{Bio}	133
C.3	The Conceptual Definitions D_{Bio}	147

1 Introduction

The probably most important source of biochemical data is the fast growing number of articles available in electronic form. Medline for example contains over 10 million abstracts and approximately 40.000 are added each month. Other important resources are the Journal of Biological Chemistry with more than 50.000 pages published per year and the SWISSPROT database (120.350¹ protein sequence entries), which also contains natural language texts describing the function of proteins. This huge amount of unstructured information has in fact become to be known as the “biobibliome”. Indeed it seems crucial to exploit natural language processing techniques to extract information from these free text sources and feed databases with them. The storage and organization of this biochemical knowledge in a database can in turn facilitate the reasoning about the data and lead to the understanding of specific biochemical processes as well as to the discovery of new aspects of them.

Information Extraction (IE) is the task of identifying, collecting and normalizing relevant information from natural language texts and producing a set of target knowledge structures as output. These target knowledge structures are defined by a given ontology which represents a model of the domain in question and thus also specifies which information is relevant ([40]). In fact, a lot of research in IE is concentrating on biomedical or biochemical articles as domains of application ([6], [67], [16], [63] [62], [55], [61] [74]). In particular, some researchers have focused on the extraction of events, i.e. the dynamic aspects of the domain in question ([62], [55], [74]). Most state-of-the-art information extraction systems are limited to the extraction of isolated events without situating them properly within the context of other extracted events. However, the following two examples taken from the SWISSPROT database clearly show the necessity to establish contextual dependencies between events:

- (1) BINDS STEM LOOP II OF U1 SNRNA. [...] THIS INTERACTION IS REQUIRED FOR THE SUBSEQUENT BINDING OF U2 SN-RNP AND THE U4/U6/U5 TRI-SN-RNP.
- (2) THIS PROTEIN BINDS THE HIV-1 TATA ELEMENT AND INHIBITS TRANSCRIPTIONAL ACTIVATION BY THE TATA-BINDING PROTEIN (TBP).

In the first example, it is important to resolve the definite description ‘THIS INTERACTION’ as referring to the binding event mentioned in the first sentence. Only then will we get the correct interpretation that the binding event of the first sentence is the one ‘REQUIRED FOR THE SUBSEQUENT BINDING’ mentioned in the second one.

In the second example, it is clearly not enough to extract the *bind* and *inhibit* events in isolation. Only if we identify that the relation between the extracted events is a resultative one, will we yield the correct interpretation of the sentence, i.e. that it is the binding of TMF to the HIV-1 TATA element which

¹This number corresponds to the state of 06-Dec-2002.

inhibits the transcriptional activation by TBP.

It has become clear that it is not enough to extract isolated events but that they have to be embedded within the context they are extracted from. Thus on the one hand the necessity of a linguistic approach which identifies conceptual relations between extracted events seems obvious. On the other hand information extraction systems are typically restricted to a specific domain of application which makes it possible to create a conceptual model of the domain in question and exploit this domain knowledge within such an approach.

It is also interesting to notice that (1) is an example of the famous bridging phenomenon ([10]). Within this work, *bridging* is understood in line with Asher et al. ([2]) as the phenomenon that a linguistic expression introduced in a text refers to a certain antecedent and that they are related in a way which is not explicitly stated. In this sense, if a *bridging reference*, i.e. the relation between the two expressions, is not resolved, the text will become incoherent. Here follow some examples ² taken from Clark ([10]):

- (3) I walked into the room. <The chandelier(s)> sparked brightly.
- (4) I met two people yesterday. <The woman> told me a story.
- (5) John was murdered yesterday. <The murderer> got away.

In all three cases, the referring definite description is related to the antecedent in a way which is not explicitly stated. The relations are *part-of*, *set-membership* and *role-of* respectively. The examples clearly show that world knowledge plays an important role in the resolution of bridging references.

Indeed example (1) can be treated and resolved as a bridging reference. In contrast, example (2) can from a theoretical point of view not be considered as such, but if verbs representing events or states are analyzed as referring expressions, relations between them and previous events can also be computed by the same approach which resolves bridging references.

The aim of the present master's thesis is to develop a knowledge-based approach to bridging reference resolution which on the basis of a given ontology and a semantic representation of events computes relations between them that are predefined in the domain model. The domain of application is molecular biology. As mentioned in [59], this is a non-trivial domain that contains most of the linguistic phenomena which typically pose problems for any NLP system. In particular, the approach will be applied to short texts describing the function of proteins within a certain organism taken from the SWISS-PROT database ([3]).

The structure of this thesis is as follows: section 2 describes the corpus used and motivates the necessity for bridging reference resolution also from a quantitative point of view. Section 3 discusses related work and outlines the general characteristics of the approach developed within this work. Section 4 presents the ontology-driven approach and section 5 explains the principles underlying

²In the examples given in this thesis, the referring expression is enclosed in brackets, while the antecedent appears underlined.

the design of the ontology. Section 7 describes the implementation of the system and section 8 discusses the results of its evaluation on a training and a test corpus. Finally, section 9 presents a conclusion and an outlook.

2 The SWISS-PROT Corpus

SWISS-PROT is an annotated protein sequence database ([3]). It is composed of sequence entries which in turn are composed of different line types each with their own format. The DE (DEscription) line for example contains general descriptive information about the sequence. In particular it gives the proposed official name as well as synonyms for the protein sequence in question. On the other hand, the CC line contains free text comments on the entry. It is further divided into different topics. The CC FUNCTION topic for example consists of natural language descriptions of the protein’s function.

A corpus has been built containing the DE line and the CC FUNCTION topic of 20189 SIWSS-PROT database entries. In the remainder of this thesis, this corpus will be referred to as “the SWISS-PROT corpus”. Furthermore, the DE line will be referred to as the *Names*-slot and the CC FUNCTION line as the *Function*-slot. The length of the *Function*-slots is between 1 and 26 sentences with an average of 1.6. The length in words ranges from 1 to 172 and is 22 on average.

A first analysis and classification of 1000 definite descriptions (DDs) of the SWISS-PROT corpus according to the taxonomy of DD uses proposed in [51] yielded the results in table 1. In the underlying taxonomy, *larger situation* uses correspond to definite descriptions which do not refer to a certain antecedent, but whose referent is uniquely determined by the context of utterance or the domain of the text the DD appears in. Typical examples of larger situation uses of a DD are *the pope* within a global context and *the president* in national contexts. *Unfamiliar* uses of DDs are those in which the definite description can not be interpreted as referring to a certain antecedent, but has to be accommodated. The *associative* use corresponds exactly to the bridging phenomenon as defined in section 1, i.e. the definite description is related to an antecedent in a way which is not explicitly stated. Finally, the *anaphoric* use includes those definite descriptions which are related to a previous antecedent by identity and this relation is explicit due to the fact that both expressions share a common head, as for example in:

- (6) John bought a car. He fell in love with <the car> at the very first moment he saw it.

type	occurrences
unfamiliar	543 (54.3%)
larger situation	311 (31.1%)
associative	107 (10.7%)
anaphoric	28 (2.8%)
doubt	11 (1.1%)
total	1000

Table 1: Analysis of 1000 DDs of the SWISS-PROT corpus

Before commenting these results it should be noted that the difference between the unfamiliar and larger situation uses is normally very subtle such that subjects have great difficulty in separating both classes ([51]). Thus the above classification of unfamiliar and larger situation uses should only be regarded as a rather vague one. In this sense, the only conclusion which can be drawn is that most of the DDs in the corpus (85.4%) represent new discourse entities in the sense of Prince ([53]) which can't be linked to an antecedent. It should be noted that this percentage is very high compared to other corpus studies on DDs ([51], [43]). This is probably due to the fact that on the one hand the corpus is very domain-specific so that much knowledge is presupposed. This could explain the high number of larger situation uses. On the other hand the function descriptions within the SWISSPROT database are relatively short (see above) thus not allowing too much intersentential relations between DDs, which could explain the high number of unfamiliar uses. Examples for unfamiliar DDs are the following ones³:

- 3BETA-HSD IS A BIFUNCTIONAL ENZYME, THAT CATALYZES THE OXIDATIVE CONVERSION OF DELTA(5)-ENE-3-BETA-HYDROXY STEROID, AND THE OXIDATIVE CONVERSION OF KETOSTEROIDS. THE 3BETA-HSD ENZYMATIC SYSTEM PLAYS A CRUCIAL ROLE IN THE BIOSYNTHESIS OF ALL CLASSES OF HORMONAL STEROIDS.
- A DEVELOPMENTALLY REGULATED PROTEIN IMPORTANT FOR MICROTUBULE FUNCTIONS. TIGHTLY ASSOCIATED WITH KAR3, MAY SERVE TO REGULATE THE CELLULAR COMPARTMENT IN WHICH KAR3 FUNCTIONS.

And here are also some examples of larger situation uses:

- A DNA FRAGMENT OF APPROXIMATELY 900 BASE PAIRS, ADJACENT TO THE FLJB (H2) GENE, WHICH SPECIFIES THE SYNTHESIS OF PHASE-2 FLAGELLIN, CAN EXIST IN EITHER ORIENTATION WITH RESPECT TO FLJB. THE ORIENTATION OF THE INVERSION REGION CONTROLS EXPRESSION OF FLJB. THE HIN GENE OCCUPIES ABOUT TWO-THIRDS OF THE INVERSION REGION; IT IS REQUIRED FOR THE INVERSION OF THE FLJB CONTROLLING REGION.
- ABI MAY INTERACT WITH A TARGET IN THE CELL MEMBRANE, WHICH COULD BE THE PRODUCT OF THE HOST'S CMRA GENE, AND CAUSE DISRUPTION OF THE CELLULAR MEMBRANE SUCH THAT LYSIS OF THE INFECTED CELL AND DEATH OF THE INFECTING PHAGE WOULD RESULT.

The interesting class of the analysis is the *bridging* one. The results show that more or less 10% of the DDs of the corpus are related to an antecedent in a way

³The complete protein function description is given for each example.

verb	occurrences	percentage of all verbal forms
INVOLVED	10675	6.2%
BINDS	6339	3.7%
CATALYZES	5346	3.1%
REQUIRED	3917	2.3%
PLAY	2856	1.7%
PLAYS	2137	1.3%
ACTS	1980	1.2%
FORM	1750	1.0%
BINDING	1354	0.8%
PROMOTES	1307	0.77%
KNOWN	1249	0.73%
INHIBITS	1236	0.73%
ACT	1124	0.66%
INTERACTS	1076	0.63%
MEDIATES	997	0.59%
COUPLED	994	0.58%
STIMULATES	935	0.55%
TRANSFERS	921	0.54%
ACTIVATED	918	0.54%
INCLUDING	888	0.52%
.	.	.
.	.	.
.	.	.
Total verbal forms	170232	

Table 2: Frequency distribution of the verbs of the SWISS-PROT corpus

which is not explicitly stated in the text.

As mentioned in the introduction, the aim of this thesis is to resolve bridging references to events as antecedents. Many events are represented by verbs so that it would be interesting to get a picture of the verbs that appear within the corpus as well as of their distribution. An analysis of the most frequent verbs of the SWISS-PROT corpus yielded the results in table 2.⁴ For the beginning, the author decided to concentrate on the analysis of binding events as expressed by the second most frequent verb *bind* and its gerund *binding* (both together constituting 4.5% of the verbal forms of the corpus) as the meaning of the most frequent verb *involve* is too dependent on what something is involved in and thus it is more difficult to decide whether a certain expression can be understood as a bridging reference to it or not. Furthermore it is not clear if the verb *involve*

⁴Part-of-speech tagging information has been used to obtain these results; modal verbs such as *may/could/should* as well as the verb *'have'* and the copula verb *'be'* together with their derived verbal forms have been omitted.

type	occurrences
event coreference	27 (5.1%)
event inference	137 (25.9%)
event role	28 (5.3%)
total binding events	528 (100%)

Table 3: Results of the classification of binding events as antecedents

has in fact an event reading. From the author’s point of view it denotes rather a state than an event.

So all the entries from the SWISS-PROT corpus containing the verbs *bind*, *binds* and *binding* have been selected. Out of the resulting 3623 entries, 500 have been randomly chosen. A detailed study of this entries allowed to distinguish three relationships between events as antecedents and a definite description or some other event or state as referring expression:

- as *event coreference* will be regarded those linguistic expressions referring to one and the same event, as in the following example:
(ROB) BINDS TO THE RIGHT ARM OF THE REPLICATION ORIGIN ORIC OF THE E.COLI CHROMOSOME. <ROB BINDING> MAY INFLUENCE THE FORMATION OF THE NUCLEOPROTEIN STRUCTURE, REQUIRED FOR ORIC FUNCTION IN THE INITIATION OF REPLICATION.
- as *event inference* will be understood those cases in which domain knowledge is necessary to establish the relation between an expression and an event, as for example in:
THIS PROTEIN BINDS THE HIV-1 TATA ELEMENT AND <INHIBITS> TRANSCRIPTIONAL ACTIVATION BY THE TATA-BINDING PROTEIN (TBP).
- those bridging descriptions in which information about a certain event is added by further describing a (possibly) implicit role of it as in the following example will be considered as *event role*:
(GPAR alpha) BINDS TO CNTF (GPA). <THE ALPHA CHAIN> PROVIDES THE RECEPTOR SPECIFICITY.

The author has classified the binding events of the 500 entries mentioned above into the three suggested categories. The results are summarized in table 3 and show that well above one third of the binding events in the corpus represent an antecedent for some other expression. Thus the necessity of resolving bridging references to events as antecedents in order to establish discourse coherence becomes also clear from a quantitative point of view.

3 Related Work

3.1 Bridging

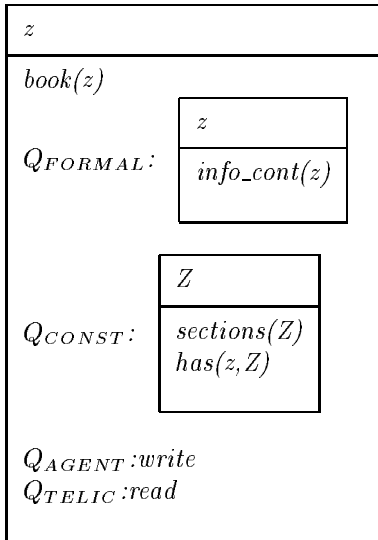
The author is aware of four formal accounts of the bridging phenomenon he would like to discuss within this section. These are the 'Bridging as Coercive Accommodation' approach of Bos et al. ([7]), the 'Bridging as Byproduct of Discourse Interpretation' analysis of Asher and Lascarides ([2]), the 'Interpretation as Abduction' framework of Hobbs et al. ([22]) as well as Piwek et al.'s Constructive Type Theory approach ([48]). Furthermore, there also exist three empirically motivated approaches which will be briefly described at the end of the section. Concerning discourse analysis in information extraction, three systems which took part on the MUC Coreference Task are discussed in detail. The section concludes with an overview of state-of-the-art ontologies for molecular biology which in some way or the other have influenced the work presented within this thesis.

3.1.1 Bridging as Coercive Accommodation

Bos et al. center their work on bridging resolution around definite descriptions as presupposition triggers. They follow the line of Van der Sandt's theory of presupposition projection thus taking presuppositions to behave as pronominal anaphora to be linked to previously established antecedents. However, bridging resolution differs from anaphora resolution in that the antecedent is not explicitly given and must be implicitly inferred from the context. To account for this, Bos et al. make use of a *qualia structure* for each lexical element. The idea of a *qualia structure* goes back to Pustejovsky ([54]) and represents a way of encoding world knowledge for each lexical element. A qualia structure entry basically contains four slots corresponding to the four qualia roles:

- FORMAL (the formal definition of an object)
- CONSTITUTIVE (the relation between an object and its parts)
- TELIC (the purpose and function of the object)
- AGENTIVE (the factors involved in the 'creation' the object)

The DRT-style lexicon entry for *book* together with its qualia structure for example looks as follows:



It should be mentioned that the qualia-information represented in Q-DRSs is not accessible (see [26] for a definition of DRT-accessibility) and does not affect the truth-conditions of a DRS. It is introduced in the lexicon and brought into discourse via the DRS bottom-up construction algorithm. If necessary, for example to play the role of an antecedent, the qualia structure can be “raised” to the surface by a process Bos et al. define as *coercive accommodation*:

Definition 1 (Coercive Accommodation (CA))

$CA(\langle U, C \rangle) = \{\langle U, C \rangle \oplus K | Q : K \in C\}$, where \oplus is the merging operator defined on DRSs, U and C respectively the set of discourse referents and the conditions of a DRS ([26]).

Following Bos et al., anaphora resolution takes place in three different ways:

1. resolution to an accessible, suitable discourse referent (*linking*)
2. resolution to coercively accommodated material of an accessible DRS (*bridging*)
3. accommodation of the anaphoric information to an accessible DRS (*accommodation*)

Suitability is understood by Bos et al. as an extra constraint on the choice of an antecedent. Informally, a DRS K_2 is suitable to another DRS K_1 if there is a way to map the discourse referents $U(K_2)$ of K_2 to the discourse referents $U(K_3)$ of a DRS K_3 and in addition $C(K_3) \subseteq C(K_1)$ (where \subseteq is the subordination relation between DRSs defined in [26]).⁵

The Anaphora Resolution (AR) algorithm works by linking all the anaphoric

⁵See section 4 for a formal definition of suitability. The definition of Bos et al. is not given here as it differs from the one underlying this work and thus to avoid confusion.

expressions to a suitable antecedent and thus resolving them or by accommodating them thus introducing a new discourse referent. Anaphoric expressions are marked by α -DRSs and only after having been merged with the discourse processed so far resolved. The resolution process then makes the α -marks disappear so that afterwards we yield proper DRSs which are model-theoretically interpretable as in standard DRT ([26]).

The following formal definitions of the Anaphora Resolution process make clear that linking is preferred to bridging and bridging is preferred to accommodation. So accommodation is regarded as the emergency case when everything else fails. Note that K_α represents an anaphoric DRS, K_m represents the main DRS and $K_1[K_2/K_3]$ means that K_2 is substituted by K_3 in K_1 :

Definition 2 (Anaphora Resolution (AR))

$$\begin{aligned} AR(K_\alpha, K_m) &= \{K' | K' \in LINK(K_\alpha, K_m)\} \text{ iff } |LINK(K_\alpha, K_m)| > 0 \\ &= \{K' | K' \in BRIDGE(K_\alpha, K_m)\} \text{ iff } |LINK(K_\alpha, K_m)| = 0 \text{ and} \\ &\quad |BRIDGE(K_\alpha, K_m)| > 0 \\ &= \{K' | K' \in ACC(K_\alpha, K_m)\} \text{ iff } |LINK(K_\alpha, K_m)| = 0 \text{ and} \\ &\quad |BRIDGE(K_\alpha, K_m)| = 0 \end{aligned}$$

Here follow the formal definitions of the *linking*, *bridging* and *accommodation* operations (taken slightly modified from [7]):

Definition 3 (Linking)

$$\begin{aligned} LINK(K_\alpha, K_m) &= \{K'_m | K_2 \leq K_1 \leq K_m \wedge \alpha : K_\alpha \in C(K_2) \wedge \\ &\quad K_\alpha \text{ m-suitable to } K_1 \wedge \\ &\quad U(K_3) = U(K_2) \cup U(K_\alpha) \wedge \\ &\quad C(K_3) = C(K_2) - \alpha : K_\alpha \cup C(K_\alpha) \cup \{x = y | m(x) = y\}\} \\ &\quad K'_m = K_m[K_2/K_3] \end{aligned}$$

Definition 4 (Bridging)

$$\begin{aligned} BRIDGE(K_\alpha, K_m) &= \{K'_m | K_2 \leq K_1 \leq K_m \wedge \alpha : K_\alpha \in C(K_2) \wedge \\ &\quad K_{CA} \in CA(K_1) \wedge \\ &\quad K_\alpha \text{ m-suitable to } K_{CA} \wedge \\ &\quad U(K_3) = U(K_2) \cup U(K_\alpha) \wedge \\ &\quad C(K_3) = C(K_2) - \alpha : K_\alpha \cup C(K_\alpha) \cup \{x = y | m(x) = y\} \wedge \\ &\quad K'_m = K_m[K_2/K_3][K_1/K_1 \oplus K_{CA}]\} \end{aligned}$$

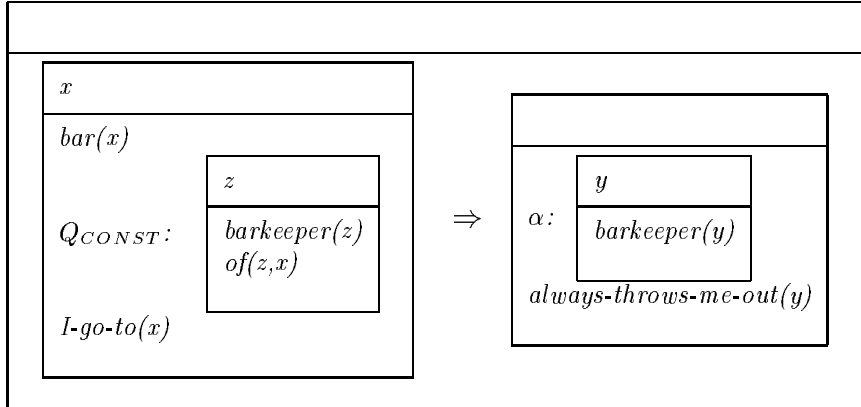
Definition 5 (Accommodation)

$$\begin{aligned} ACC(K_\alpha, K_m) &= \{K'_m | K_2 \leq K_1 \leq K_m \wedge \alpha : K_\alpha \in C(K_2) \wedge \\ &\quad U(K_3) = U(K_2) \\ &\quad C(K_3) = C(K_2) - \alpha : K_\alpha \\ &\quad K'_m = K_m[K_1/K_1 \oplus K_\alpha][K_2/K_3] \end{aligned}$$

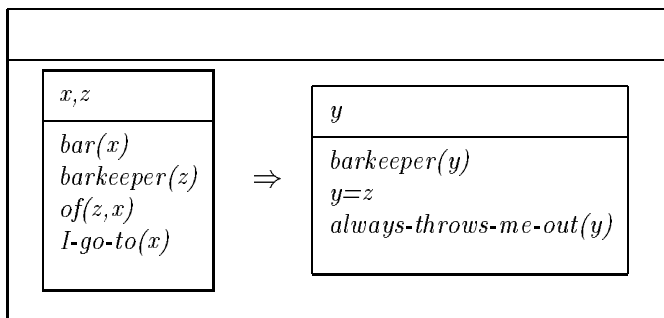
Bos et al's approach can be illustrated on the following example (from [7]):

- (7) When I go to a bar, the <barkeeper> always throws me out.

The DRS for this example looks as follows:



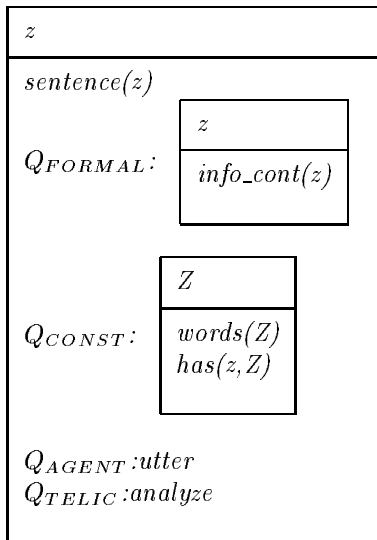
In the example bridging is successful because the presupposition trigger *the barkeeper* can be linked to an element of the qualia structure of the *bar* thus yielding the following resolved DRS in which the inferred barkeeper has been coercively accommodated:



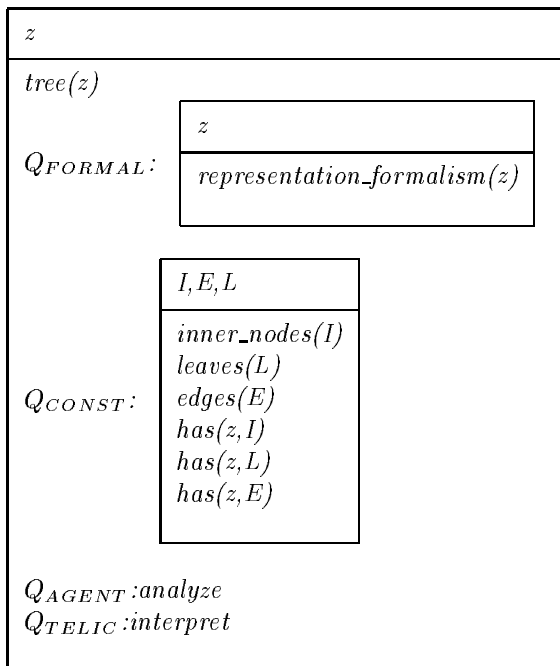
However, though their approach successfully accounts for examples where lexical knowledge is enough to drive the bridging reference resolution (as in example (7)), it will certainly fail for examples where more complex world knowledge is necessary. Piwek et al. ([48]) for example have argued that Bos et al.'s approach will fail to explain the following example:

- (8) Yesterday Chomsky analyzed a sentence on the blackboard, but I couldn't see <the tree>.

The necessary knowledge to resolve the bridging reference in (8) is that 1.) sentences are analyzed by means of a representation formalism and that 2.) a tree is a type of a representation formalism. A qualia structure for *sentence* in the sense of Bos et al. would merely indicate that a *sentence* (as a book) is an information container, is formed by words, is uttered or written and can be analyzed. So the qualia structure entry for *sentence* could look as follows:



Similarly (if we understand a tree merely as a representation formalism), the qualia structure of *tree* could define that it is a representation formalism, that it has inner nodes, leaves as well as edges connecting the nodes, that it is brought about by analyzing (a sentence) and that its purpose is to be (semantically) interpreted:



So the necessary world knowledge to resolve example (8) seems to be explic-

itly available and thus we wonder why Bos et al.’s approach fails in explaining it. This is clearly due to the following reasons:

1. The respective qualia structures are unrelated between each other, such that there is no way to relate the $Q_{TELIC} : \textit{analyse of sentence}$ with the $Q_{AGENT} : \textit{analyse of tree}$.
2. The qualia structure basically represents lexical knowledge where a qualia structure for an event like $\textit{analyze}(e, \textit{'Chomsky'}, s)$ would seem more likely to provide all the necessary knowledge for the correct linking of the definite description *the tree* to the event e .

3.1.2 Bridging as Byproduct of Discourse Interpretation

Asher and Lascarides ([2]) also focus on definite descriptions within their approach to bridging reference resolution. However, their approach differs substantially from the one of Bos et al.

Their main claim is that “bridging is a byproduct of discourse interpretation”. And more precisely that “Bridging inferences are a byproduct of computing how the current sentence connects to the previous ones in the discourse.”

They formalize their approach within a theory known as SDRT which is an extension of DRT ([26]). In SDRT, a discourse is represented by segmented DRSs (SDRTs) which are linked together with rhetorical relations such as *Narration* or *Parallel* ([34],[35]).

These relations are inferred via a ‘glue’ logic called DICE which is based on a nonmonotonic notion of validity (\approx) and exploits the use of defeasible rules making use of a weak conditional operator $>$. $P > Q$ in this sense means something like: if P then normally Q .

The DICE rule for *Narration*, for example, states that if a discourse segment β is to be attached to α with a rhetorical relation, where α is part of the discourse structure τ already constructed (i.e. $\langle \tau, \alpha, \beta \rangle$ holds) and α and β describe events (as opposed to states) (i.e. $\textit{event}(e_\alpha)$ and $\textit{event}(e_\beta)$ hold), then normally the rhetorical relation is *Narration*, or more formally:

Definition 6 (Narration)

$$\langle \tau, \alpha, \beta \rangle \wedge \textit{event}(e_\alpha) \wedge \textit{event}(e_\beta) > \textit{Narration}(\alpha, \beta)$$

The *Temporal Consequence of Narration* is a coherence constraint on *Narration* which states that if α and β are related by *Narration* then α ’s event precedes β ’s:

Definition 7 (Temporal Consequence of Narration)

$$\textit{Narration}(\alpha, \beta) \rightarrow e_\alpha \prec e_\beta$$

The *Spatial Consequence of Narration* is also a coherence constraint on the semantic relation between the constituents connected by *Narration*. It states that if the actor of e_α and e_β is the same, then his/her location must be the same at the end of event e_α as at the beginning of event e_β :

Definition 8 (Spatial Consequence of Narration)

$$(Narration(\alpha, \beta) \wedge actor(x, \alpha) \wedge actor(x, \beta)) \rightarrow loc(x, goal(e_\beta)) = loc(x, source(e_\alpha))$$

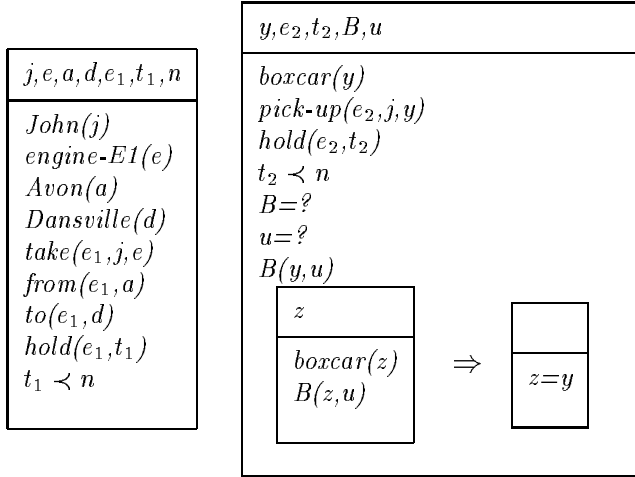
Asher and Lascarides define their approach to bridging resolution via four axioms specifying how the content of the discourse τ is to be updated.

The first of the four axioms mentioned (*If Possible use Identity*) corresponds to the preference for linking to bridging in Bos et al’s approach and captures Van der Sandt’s intuition that identity should be used to resolve bridging references if possible. The second axiom (*Bridges are Plausible*) states that bridging reference resolutions must be plausible. The third one (*Discourse Structure (DS) Determines Bridging*) captures the intuition that if a certain rhetorical relation has been inferred via DICE and this rhetorical relation specifies a way of resolving the bridging reference, then it will be done that way. The last rule (*Maximize Discourse Coherence*) states that if there are several plausible ways to resolve a bridging reference, one should take the one which allows to infer the rhetorical relation leading to the highest discourse coherence. This obviously presupposes an order $>_r$ defined on the rhetorical relations.

In fact, Asher’s and Lascarides’ approach gives a satisfactory account of the bridging phenomenon in the following example:

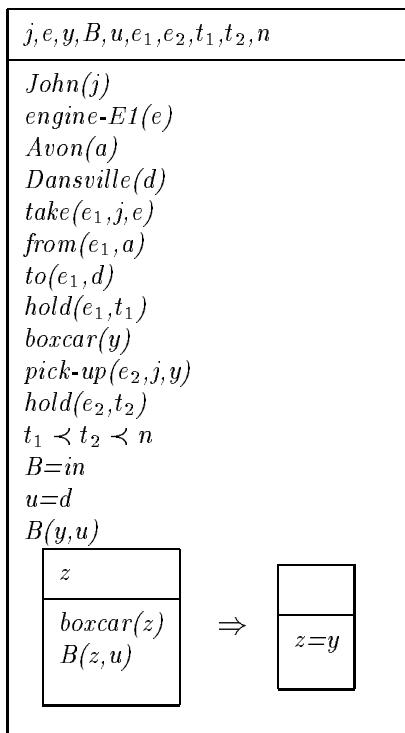
- (9) John took engine E1 from Avon to Dansville. He picked up the boxcar and took it to Boxburn.

The corresponding DRSs for the two sentences are respectively (assuming that the pronoun ‘He’ is already resolved):



It should be first mentioned that the analysis of definite descriptions underlying Asher and Lascarides’ work is that they are related to a certain antecedent u in a specific way B and furthermore that the Russelian uniqueness condition holds ([2]). As both sentences represent events, the rhetorical relation *Narration* is nonmonotonically inferred (see definition 6). Furthermore, because of

the semantics of the verbs *take to* and *pick up*, which should define that the agent of the action is supposed to be at the same location as the *taken* or *picked up* object, and the *Spatial Consequence on Narration*, one infers that in (9) the source of the *picking up* event is in Dansville and thus that *the boxcar* is also there. Thus, the coherence constraints on *Narration* allow us to resolve B to *in* and u to *Dansville*:



However, as Asher and Lascarides also mention, there is not always enough information within the discourse to infer a particular rhetorical relation. Thus, within their 'bridging as byproduct of discourse interpretation' theory, the *Maximize Discourse Coherence* axiom has to be understood as an emergency rule: when no rhetorical relation can be inferred, it is the resolution itself which leads to a certain rhetorical relation and not the other way round. Let's discuss the following example (out of [2]):

(10) I just arrived. The camel is outside and needs water.

In the author's opinion, Asher's and Lascarides' treatment of this example is very poor. They correctly argue that *the camel* can't be linked to the previous discourse by identity and that the relation *Background* can't be inferred because of lack of a common topic. Furthermore, they claim:

[...] we must entertain various resolutions of the underspecified parameters in β and see which option maximizes discourse coherence.

Suppose B and u are resolved so that the camel had some role in the arrival. By the constraint *Bridges are Plausible* this must be a plausible role. The only one is that the camel is the mode of transport by which I arrived. This content enables us to infer a new rhetorical relation, with improved discourse coherence. We can infer that the camel being outside was caused by my arrival thanks to the spatial information in the compositional semantics of the change of location phrase *arrive here*, and so the rhetorical relation is Result. So *Maximize Discourse Coherence* is used to infer this new content to the definite description *the camel*, together with the *Result* relation between the constituents.

The problem of Asher’s and Lascarides’ approach is that they don’t explain what they mean with the fact that “a certain resolution is plausible” and don’t spell out the resolution process itself. Thus within their approach, in the case that no discourse relation can be inferred, resolution seems like an arbitrary search in the space of all possible relations B and all possible antecedents u constrained by a plausibility criterion which they don’t define at all.

3.1.3 Interpretation as Abduction

In Hobbs et al.’s ‘Interpretation as Abduction’ approach, sentences are interpreted by proving or deriving their logical form from a knowledge base constituted by world knowledge as well as the information content of the discourse processed so far. Within their approach, weighted abduction is used as inference mechanism. In brief, abduction means that from $(\forall x)p(x) \supset q(x)$ and $q(A)$ we conclude $p(A)$. The abduction is weighted in the sense that each construct gets assigned a weight representing the cost of assuming its existence or - in Lewis’ terms ([39]) - accommodating it. The goal of this abduction mechanism is to get an interpretation, i.e. a proof of the sentence, at the minimum cost. Thus in particular bridging references are resolved by finding a (sub-) proof for them. Their work is related to this one in the use of world knowledge in form of axioms and the use of an inference mechanism to drive the bridging resolution process. However, it is rather unclear how to (relatively) set the weights for accommodating different constructs. Furthermore, as world knowledge and the information content of the discourse so far are mixed together in one knowledge base from which a proof for the sentence is derived, it is not intuitive how to separate the direct anaphora and the bridging classes.

3.1.4 Constructing Bridges

Piwek et al. ([48]) rephrase Van der Sandt’s approach to presupposition projection in terms of Constructive Type Theory (CTT). Their approach is closely related to Hobbs et al.’s in the sense that they also exploit world knowledge to drive the bridging resolution process by using a proof system. However, the

advantage of using a formalism like CTT is not clear to the author, especially because there are proof systems for DRT such as the DRS calculus of Kamp and Reyle ([27]) and furthermore because DRT has proved so valuable for discourse representation and the analysis of discourse phenomena such as pronoun resolution ([26]), presupposition projection ([69]), bridging ([7]), topic-comment structures ([73]), temporal underspecification ([60]) etc.

3.1.5 Empirically Based Approaches

The motivation underlying Poesio et al.’s approach to bridging reference resolution ([70], [72], [52], [51], [71], [49]) is a corpus-based analysis of definite descriptions which shows that more or less 50% of them are non-anaphoric (*unfamiliar* or *discourse new* respectively in Hawkins’ ([21]) and Prince’s ([53]) terms). The conclusion drawn by Poesio et al. is that systems attempting to resolve definite descriptions should classify them as anaphoric or non-anaphoric before actually trying to link them to a certain antecedent. Poesio et al. make use of lexically and syntactically based heuristics for this purpose. The algorithm used by them is basically as follows:

1. use ‘safe’ heuristics to classify a definite description as non-anaphoric
2. if that fails, try to find an antecedent with the same head
3. if that also fails, apply other heuristics to identify the definite description as *discourse new*
4. only then try to interpret the definite description as a bridging reference

In Poesio et al.’s approach WordNet is used as source of lexical knowledge driving the resolution of bridging references ([52], [72]) but recently also methods for acquiring lexical information such as hyponymy or meronymy relations from corpora have been explored ([49]). The performance of Poesio et al.’s system on Wall Street Journal articles was a recall of 57%, a precision of 70% and thus an *F-measure* of 62%. The recall (R) is a measure of how many of all the possible correct answers are found by the approach, while the precision (P) is a measure of how many of the total answers given are actually correct:

$$R = \frac{\# \text{correct answers given}}{\# \text{total correct answers}} \quad (11)$$

$$P = \frac{\# \text{correct answers given}}{\# \text{total answers given}} \quad (12)$$

The *F-measure* is a metric which combines recall and precision into a single value using the formula:

$$F = \frac{(\beta^2 + 1.0) * P * R}{\beta^2 * P + R} \quad (13)$$

where β is the relative weight given to recall over precision. Within the work presented here all F-measures will be calculated using $\beta = 1.0$, i.e. giving equal weight to P and R.

Muñoz et al. ([43], [45], [44]) tackle the problem of resolving coreferential bridging references in Spanish by using the Spanish counterpart to WordNet. Their parting thoughts are similar to the ones of Poesio et al. in the sense that their algorithm also tries to rule out non-anaphoric definite descriptions at early steps in the processing in order to reduce the overall computational time. However, the search for a suitable and semantically compatible antecedent is not limited regarding the distance in words or sentences to the referring expression, which in the author's opinion is neither reasonable from a linguistic nor a computational point of view. A very interesting feature of their approach is that they also consider verbs as potential antecedents and attempt to resolve bridging references standing in a thematic role to verbal antecedents. The overall results of their anaphora resolution approach are a recall of 75.5% and a precision of 79.3% measured on data from three different corpora ([44]). It is interesting to observe that the precision of identifying a certain definite description as thematic role of a verbal antecedent is particularly low (60.9%) ([43], [44]).

Hahn et al. ([20]) present a conceptual reasoning approach to bridging reference resolution or - as they call it - textual ellipsis. Within their approach, two linguistic expressions are linked together by a specific relation if there is a corresponding path between both ontological concepts with regard to a given concept hierarchy. It is important to notice that they do not only consider identity as possible relation between two expressions but also other relations such as *has-physical-part*, *property-of* etc. Their system achieves a recall of 55.05% and a precision of 95.2% measured on a small test set of their corpus.

3.2 Discourse Analysis in Information Extraction

Discourse analysis within information extraction systems typically boils down to coreference resolution and template merging as defined in the MUC tasks. The author is particularly aware of three systems which took part in the MUC Coreference Tasks. These are the LaSIE system developed by the University of Sheffield ([24], [17], [23]), the Proteus system from New York University ([75], [18]) and the Circus system by the University Massachusetts ([64], [65], [41]). Humphreys et al. from the University of Sheffield present a knowledge-based approach to coreference resolution making use of an explicit semantic representation in form of a predicate-argument structure ([17], [23]). The input to their coreference resolution algorithm are sets of pairs of potentially coreferring entities for which the referring entity is always the same. The algorithm is defined as follows:

1. ensure semantic type consistency
2. ensure non-distinctness
3. ensure attribute consistency

4. calculate similarity score
5. choose pair with highest similarity score

In the first step the semantic type compatibility of the entities in question is ensured. In this sense two entities are compatible if their corresponding types are ordered with regard to a given taxonomy. The second step ensures that there do not exist any constraints which could hinder the two entities to be resolved as coreferring. This is for example the case if the referring expression is an indefinite. The third step verifies that the value of the common attributes of the corresponding predicate-attribute structures are identical. Then the similarity score is calculated as the sum of a semantic similarity score and an attribute similarity score, where the semantic similarity score is inversely proportional to the length (in nodes) of the taxonomic path between the two types and the attribute similarity score is the ratio of attributes with compatible values against the total number of attributes. Finally, the highest similarity score of the whole set then decides which entities are resolved as coreferring. The results of this algorithm on the entity coreference task were a recall of R=50.71% and a precision of P=71.93% on the MUC-6 management succession scenario and P=56.1% and R=68.8% on the MUC-7 launch event scenario ([24]). In [23], the above algorithm has also been applied to event coreference resolution. However, Humphreys et al. do not give any explicit quantitative results, but report a slight overall improvement of their whole information extraction system on the management succession task when adding the event coreference resolution module.

The discourse component of the PROTEUS system ([75], [18]) is basically similar to the one used in LaSIE. Grishman et al. also make use of an explicit semantic representation they call 'logical form' and identify two entities as coreferring if the class of the referring entity is equal or more general than the one of the antecedent with regard to a given taxonomy and the arguments of both logical forms can be matched. The results of their entity coreference algorithm on the MUC-6 management succession scenario were a recall of R=53% and a precision of P=62% ([18]).

Lehnert et al. from the University of Massachusetts make use of a machine learning approach ([64], [65], [41]) to entity coreference resolution. In particular, they use decision trees trained on hand-coded target output to classify pairs of entities (nouns, noun phrases and pronouns) as coreferring or not. Though their approach shows very good results on the MUC-5 English Joint Venture corpus ([41]), as these results are highly dependent on the set of features used, it is not clear if the same set of features and the same training parameters would also produce comparable results on other corpora. In fact, as mentioned in [41] the choice of the appropriate features for training the decision trees is still a topic of ongoing research. Furthermore, the fact that the domain knowledge for the discourse analysis step is not available in a declarative form, but is encoded in decision trees, hinders this knowledge from being exploited by other systems performing on the same or a similar domain and even by other components of the same system.

3.3 Ontologies for Molecular Biology

Most of the work concerning the development of ontologies for molecular biology does not go beyond the modeling of taxonomic or mereological relations between concepts. The GENIA ontology for example ([46]) basically provides a taxonomy of chemical objects (sources, chemical substances, etc.). [47] presents a UML-based model of static relations between biochemical objects like for example that a genome contains several chromosomes, that a chromosome in turn contains several chromosome fragments, etc. The Gene Ontology ([12], [13]) goes one step further and also includes processes such as DNA unwinding or DNA strand elongation within its taxonomy. However, it does not establish any conceptual links between the different processes. Furthermore, the edges between different nodes of the directed acyclic graph representing the 'ontology' would in the author's opinion benefit from a clear semantics as sometimes it is not clear if an edge represents a *is-a* or a *part-of* relation. [15] represents an interesting approach to modeling signal transduction pathways in form of a graph but also shows a lack of a clear semantics of its edges.

The EcoCyc ontology ([29]) on the other hand presents an interesting ontology of biochemical functions and reactions making use of Frame knowledge representation systems (FRSs) as declarative representation formalism. However, the most interesting work concerning the modeling of biochemical pathways the author is aware of is [63]. Rzhetsky et al. ([63]) do not only provide a taxonomy of biochemical objects, biochemical interactions, reactions and processes but also model conceptual relations between them. Furthermore, they distinguish between logical and biochemical concepts. In this sense, the fact that A activates B has to be considered as a logical representation of a biochemical process such as the phosphorylation, dephosphorylation, cleavage or binding of B by A. In line with [66], they also give axiomatic definitions of the relations which are relevant for the modeling of regulatory networks. For example the *activation* relation is defined as being transitive, i.e. if A activates B and B activates C then it is also valid to state that A activates C.

3.4 Discussion

In ([11]) Clark proposes a model of definite reference use in which speakers are allowed to use a definite reference if the event or object they refer to is in the speaker's and the addressee's memory or if it is part of the knowledge shared within the community they both belong to.⁶ When applying these ideas to written language, the text would play the role of the speaker, the discourse processed so far that of the memory, the domain knowledge that of the shared community knowledge and the reader or NLP system that of the addressee. Thus to understand and resolve definite or bridging references within written texts we also need on the one hand a 'memory', i.e. a formalism making the linguistic

⁶The notion of mutual knowledge plays a central role in the explanation of the conditions which license a speaker to make use of a definite reference. The interested reader is referred to [11].

context explicit, as well as a declarative formalism to represent knowledge about the domain the text deals with.

The first criterion, i.e. the need to represent the linguistic context or discourse will be satisfied in an elegant way by using DRT. DRT is definitely suitable for this purpose as it makes entities introduced in a text explicit as discourse referents that can be referred to later in the discourse ([26]). Furthermore, DRT's notion of accessibility imposes structural constraints on the choice of a certain antecedent. In particular it explains why in the following minimal pair, the definite description *the roses* can only be resolved as a bridging reference to the *bunch of flowers* in the first example:

(14) John gave Jenny a bunch of flowers. She liked <the roses> very much.

(15) John didn't give Jenny a bunch of flowers. She liked the roses very much.

Having said how the linguistic context will be represented, it remains to be explained how to represent domain knowledge. Considering example (3), it is clear that the resolution of the definite description 'the chandeliers' and its linking to the room by the relation *part-of* is only possible if we have the world knowledge that rooms have lamps and that a chandelier is some sort of lamp. Generally speaking, we need definitions of entities or concepts specifying their constituents or other entities related to them in some way. But we also need taxonomic relations like the one that a chandelier is a type of lamp. Thus it seems that bridging reference resolution involves knowledge in form of definitions of concepts as well as taxonomic relations, and this is the reason why in the author's opinion an ontology is so suitable as declarative representation of domain knowledge.

The approach to bridging reference resolution presented in this thesis is basically a combination of Bos. et al's and Asher et al's approach. Certainly both approaches are very different in their nature. The driving force behind Bos et al's approach is the use of a *qualia structure*, which clearly has to be identified as an attempt to make world knowledge explicit and use this knowledge to provide the necessary inferences for the bridging resolution process to be successful. The driving force behind Asher's and Lascarides' approach is the DICE-driven computation of rhetorical relations together with the respective coherence constraints derived from them. These coherence constraints have to be understood as a form of (rather linguistically motivated) world knowledge capable of driving and constraining the resolution process. However, the two approaches have a few things in common the author would like to point out:

1. the use of DRT as framework to formalize their respective approaches
2. the use of world knowledge to drive the resolution process
3. the restriction to definite description as presupposition triggers
4. the explicit (though different) marking of expressions which need to be resolved, i.e. projected

5. a preference to resolve definite descriptions by identity

As mentioned above, the approach described here will also make use of DRT to represent the discourse. Furthermore, world knowledge will be represented in form of an ontology. The obvious advantage of an ontology (in comparison to a qualia structure like the one of Bos et al.) is that objects can be hierarchically (taxonomically) ordered and related to each other. Furthermore, the objects represented in this ontology will not just be lexical items as in Bos et al's approach, but events and states, which in the author's opinion allow a more accurate and complete specification of knowledge than lexical items alone do.

In the approach presented in this thesis, events and their semantic representation will play a central role in the bridging resolution process. In fact, as mentioned in section 1, the approach described in this thesis will focus on events as antecedents with the intention to account for those classes of bridging references which Clark called: *Necessary roles*, *Optional roles* and *Reasons, causes, consequences and concurrences*. Most of the work on bridging reference resolution has concentrated on definite descriptions as presupposition triggers. The approach described here will follow this line but additionally consider verbs representing events and states as presupposition triggers in the sense that they are related to previous events in some way thus establishing discourse coherence. This is completely in line with Clark's definition of the *Reasons, causes, consequences and concurrences* class:

The Antecedent to the Given information of a sentence is often an event and not an object, and then it plays different types of roles with respect to previous events. Instead of being agents, objects, or instruments characterized with respect to previously mentioned events, this class of Antecedents give reasons for, causes of, consequences to, or concurrences of previously mentioned events or states.

The author would like to point out that definite descriptions do not always represent entities, but also events and states. The distinction between these three types becomes important when identifying the relation between the referring expression and the antecedent and poses constraints on its choice. Between an event and a referring entity only the relation *role-of*⁷ will be possible. The relations considered between events and states are Lascarides et Asher's discourse relations ([35]) as well as the identity relation between two events (coreference). So rhetorical relations will also play a role in the approach but as a byproduct of bridging and not as the driving force behind it as in [2]. Thus, the author's approach is in line with Asher's and Lascarides' *Maximize Discourse Coherence*, i.e. first bridging references will be resolved and on the basis of this resolution rhetorical relations will be inferred.

Thus, in contrast to the approaches discussed in section 3.1.5 and 3.2, the one presented here will not only focus on the resolution of coreferences between

⁷The *role-of*-relation has to be seen as a generic one which will be instantiated with more specific ones such as the *instrument*-relation in example (8).

events, but also of other conceptual relations between them such as *Result*, *Explanation* or *Elaboration*. Furthermore, in contrast to [41], it will not represent another machine-learning based approach to bridging reference resolution, but a knowledge-based approach which exploits knowledge coded in a declarative manner in form of an ontology.

4 The Ontology Driven Approach

Bos et al.'s approach as well as the one presented here follow the line of Van der Sandt's theory of presupposition projection ([69]) thus considering presuppositions to behave as pronominal anaphora to be linked to previously established antecedents. In Van der Sandt's view, anaphoric expressions can either be linked to a suitable antecedent or accommodated. However, bridging references differ from direct anaphora in that the antecedent is not explicitly given and must be implicitly inferred from the context. This is where world knowledge comes into play. As already mentioned, an ontology will be used as knowledge representation formalism and as the backbone of the inference mechanism required for the bridging resolution process.

Before formally defining how world knowledge in general and an ontology in particular can be integrated into the bridging resolution process, it should be mentioned that the author agrees with Van der Sandt and Bos et al. in the way anaphora resolution works. Unresolved anaphoric expressions will also be represented by α -marked DRSs (compare section 3). A new DRS containing α -marked DRSs will be merged with the main-DRS for the discourse processed so far as in standard DRT and only resolved after merging.⁸ As in Bos et al.'s approach, resolution can take place in three ways (*linking, bridging, accommodation*) and linking will be preferred to bridging and bridging to accommodation. When an α -marked DRS is resolved, the α -mark will disappear. Only when the main-DRS does not contain any unresolved DRSs, will it be interpretable as in standard DRT ([26]). Intuitively, a DRS K_2 will be suitable to a DRS K_1 if the conditions of K_2 match a subset of the conditions of K_1 , or more formally:

Definition 9 (Suitability)

A DRS K_2 is m -suitable to a DRS K_1 if there is a 1-1 mapping m such that $\text{Domain}(m) = \underline{U}(K_2)$ and $\text{Range}(m) \subseteq \underline{U}(K_1)$ and K_3 is the DRS K_2 with all $x \in \underline{U}(K_2)$ substituted by $m(x)$ such that $C(K_3) \subseteq C(K_1)$, where $\underline{U}(K)$ and $C(K)$ are respectively the set of the declared discourse referents as defined in [27] and the conditions of the DRS K .

The notion of suitability defined in this way is directed in the sense that if K_2 is m -suitable to K_1 this does not necessarily imply that K_1 is also suitable to K_2 . This definition of suitability in fact accounts for partial matches in which the antecedent entails the anaphora ([32]). So in particular it accounts for the non-presupposing reading of Van der Sandt's famous example:

(16) If John has an oriental girlfriend, <his girlfriend> won't be happy.

⁸The author is aware of the fact that the deepest embedded α -marked DRS has to be processed first but will not discuss this aspect further. The interested reader is referred to Van der Sandt ([69]).

Actually, the fact that the DRS

g'
$girlfriend(g')$ $of(g',j)$

⁹ is *m*-suitable to the DRS

g,j
$girlfriend(g)$ $oriental(g)$ $of(g,j)$

accounts for the *bridging*-reading of the above example, which

due to the preference of linking over accommodation is preferred to the pre-supposing one.

4.1 The Role of World Knowledge in Bridging Reference Resolution

It has already been argued that world or domain knowledge plays a crucial role within anaphora resolution and in particular bridging reference resolution. In this section it will be discussed how world knowledge in general can drive the anaphora resolution process as defined by Bos et al. For this purpose, the definition of the *bridging*-operation will be restated in more general terms thus allowing any declarative knowledge representation formalism and inference machinery to be exploited within bridging reference resolution. This has actually to be understood as an attempt to make the approach reusable independently of the knowledge representation formalism or inference method used. In section 4.3 it will then be shown how these general notions concerning the role of world knowledge can be adapted to our specific purposes and how an ontology as formal conceptualization of a domain can be integrated into the anaphora resolution process.

Let's start our discussion with a motivating example from the SWISS-PROT corpus:

- (17) (LBP) BINDS TO THE LIPID A MOIETY OF BACTERIAL LIPOPOLYSACCHARIDES (LPS), [...]. <THE LBP/LPS COMPLEX> SEEMS TO INTERACT WITH THE CD14 RECEPTOR.

The mentioned 'LBP/LPS COMPLEX' in the second sentence has to be understood as a role in the resultative state of the binding event mentioned in the first sentence. Thus the anaphora resolution process is expected to make this resultative relation explicit and this is certainly not possible without any domain knowledge. But even assuming that the necessary domain knowledge is available, the question how it can be integrated into the anaphora resolution process in a way abstracting from the form in which this knowledge is available

⁹The pronoun *his* is assumed as already resolved to 'John'

becomes the key issue towards scalability and reusability of the approach. Certainly, to make use of world knowledge in a formalism-independent manner, the following three criteria will have to be fulfilled:

1. the need of an internal representation of world knowledge
2. the availability of some type of calculus to reason on these internal structures
3. methods to translate the knowledge representation formalism used to the internal representation

Now let's take a step back and look at what we have been doing until now. Our aim is to resolve bridging references with a DRT- based approach and we are looking for an internal and abstract representation of world knowledge. So why not making use of DRT as internal knowledge representation formalism? Certainly it is expressive enough as it is equivalent to first order logic and it can be directly integrated into the anaphora resolution algorithm. Thus we just need a calculus for DRT allowing us to perform the inferences necessary for the anaphora resolution process to be successful.

Kamp and Reyle ([27]) present such a sound and complete calculus for first order DRSs and define what it means for a DRS K'_2 to follow logically from a DRS K_0 . Intuitively, it should be clear how this works for the Generalized Modus Ponens (GMP). Informally speaking, GMP is a forward-chaining rule which adds a copy K'_2 of the right hand side K_2 of a complex condition $\boxed{K_1 \implies K_2} \in K_0$ provided that K_1 can be 'matched' with some condition $K'_1 \in K_0$. Thus the question arises how such a copy mechanism works and what it means for a DRS to be matched by some other DRS.

According to Kamp and Reyle, a copy K' of a DRS K is basically an alphabetic variant of it. The following definition taken (slightly modified) from [27] will do for the purposes here:

Definition 10 (alphabetic variant)

Let $f : \underline{U}_K \rightarrow V$ be a 1-1 function with $\text{Range}(f) \cap \text{Fr}(K) = \emptyset$, where $\text{Fr}(K)$ are those discourse referents occurring free in K . Then $f(K)$ will be that DRS K' where all the discourse referents $r \in \underline{U}_K$ are renamed by $f(r)$ and all those in $\text{Fr}(K)$ are mapped to themselves.

Concerning the matching between DRSs, Kamp and Reyle define what it means for a DRS K to be homomorphically embeddable within a DRS K' :

Definition 11 (homomorphically embeddable)

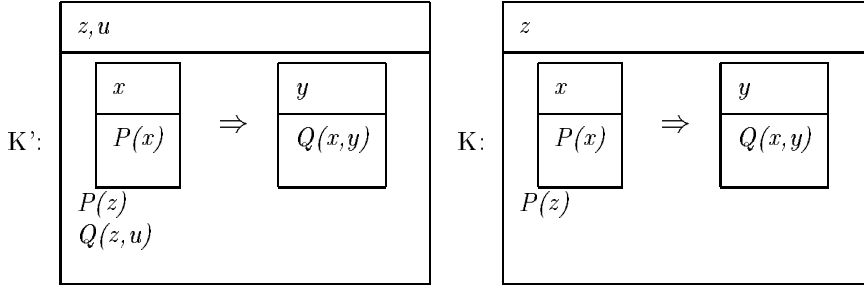
A DRS K is homomorphically embeddable in K' if there is a 1-1 function $f : \underline{U}_K \rightarrow \underline{U}'_{K'} \subseteq \underline{U}_{K'}$, such that $f(K)$ is an alphabetic variant of K and $f(K) \subseteq K'$, where the inclusion relation \subseteq between DRSs is defined according to [27].

Finally we can give a formal definition of the aforementioned forward-chaining rule (taken from [27]):

Definition 12 (DETachment or (G)eneralized (M)odus (P)onens)

Suppose $[K_1 \implies K_2] \in \text{Con}_K$ and suppose there is a homomorphical embedding f of K_1 in K . Let g be an extension of f to U_{K_2} , such that g is 1-1 and such that g maps U_{K_2} to a set of new discourse referents that are new to K . Then we may add the alphabetic variant $g(K_2)$ to K .

Let's illustrate all these definitions with a simple example. Suppose we want to prove that the DRS K' follows logically from the DRS K :

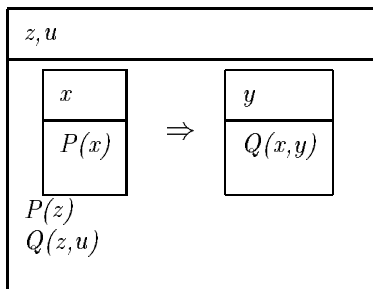


We first have to show that $\begin{array}{|c|} \hline x \\ \hline P(x) \\ \hline \end{array}$ is homomorphically embeddable in K .

This is obviously the case as $f(\begin{array}{|c|} \hline x \\ \hline P(x) \\ \hline \end{array}) = \begin{array}{|c|} \hline z \\ \hline P(z) \\ \hline \end{array}$, where $f : \{x\} \rightarrow \{z\}$

and $f(x) = z$ and in addition $\begin{array}{|c|} \hline z \\ \hline P(z) \\ \hline \end{array} \subseteq K$. Now let g be an extension of f

to U_{K_2} with $g(y) = u$. Then following our DETachment rule we can add $g(K_2)$ to K and end up with the DRS:



So far so good. But we will now perform two important modifications on Kamp and Reyle's Generalized Modus Ponens (GMP). First, the key DRS $[K_1 \implies K_2]$ allowing the forward-chaining will not be part of the DRS K from which the DRS K' follows but will be understood as forming part of our domain knowl-

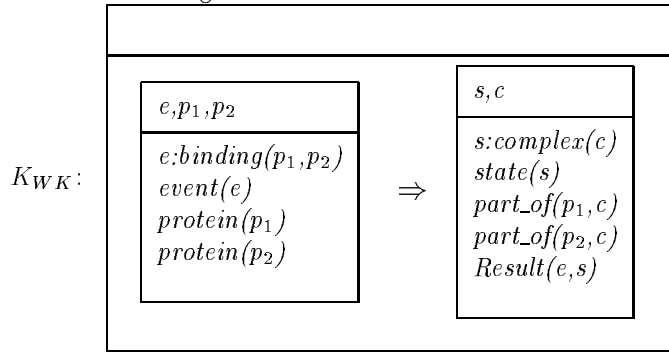
edge. Second, we will allow K and K' to be disjoint in the sense that it will not necessarily be the case that $K \subseteq K'$. Now we are ready to give a formal definition of what it means for a DRS K to imply a DRS K' with regard to a DRS $\boxed{K_1 \implies K_2}$:

Definition 13 (Implication between DRSs)

A DRS K implies a DRS K' with regard to a DRS $\boxed{K_1 \implies K_2}$ (Notation: $K \implies_{[\boxed{K_1 \implies K_2}]} K'$) iff $K \oplus \boxed{K_1 \implies K_2} \oplus K'$ follows logically from $K \oplus \boxed{K_1 \implies K_2}$ as defined by the Generalized Modus Ponens (GMP) in [27], where \oplus is the merging operator for DRSs, and it is not the case that K' follows from K alone ($K \not\subseteq K'$).

Note that according to the above definition K' is not allowed to follow trivially from K without the addition of the complex condition $\boxed{K_1 \implies K_2}$.

Let's turn to our starting example (17). Let's further assume that our domain knowledge about protein interactions contains the fact that if two proteins bind, they will form a complex and that the state of these proteins being complexed has to be interpreted as a result of this binding. This can be formalized by saying that the following complex condition K_{WK} is valid with regard to our domain knowledge:



Furthermore, assuming that the first sentence of example (17) can be represented by the DRS K (see below), it follows from definition (13) that K implies K' (see below) with regard to the complex condition K_{WK} ($K \implies_{K_{WK}} K'$) because $K \oplus K_{WK} \oplus K'$ follows logically from $K \oplus K_{WK}$ by the GMP and it is not the case that K' follows from K alone.

e', p_1', p_2'
$e': binding(p_1', p_2')$ $event(e')$ $protein(p_1')$ $protein(p_2')$ $p_1' = 'LBP'$ $p_2' = 'LPS'$

s', c'
$s': complex(c')$ $state(s')$ $part_of(p_1', c')$ $part_pf(p_2', c')$ $Result(e', s')$

Finally we will restate Bos et al's definition of the *bridging*-operation by making use of our implication between DRSs:

Definition 14 (Bridging)

$$\begin{aligned}
BRIDGE(K_\alpha, K_m) = \{ & K'_m \mid K_2 \leq K_1 \leq K_m \wedge \alpha : K_\alpha \in C(K_2) \wedge \\
& WK_L \models_L [K \Rightarrow K']_L \wedge K_1 \Rightarrow_{[K \Rightarrow K']} K'_1 \wedge \\
& K_\alpha \text{ } m\text{-suitable to } K'_1 \wedge \\
& U(K_3) = U(K_2) \cup U(K_\alpha) \wedge \\
& C(K_3) = C(K_2) - \alpha : K_\alpha \cup C(K_\alpha) \cup \{x = y \mid m(x) = y\} \wedge \\
& K'_m = K_m[K_2/K_3][K_1/K_1 \oplus K'_1]\}
\end{aligned}$$

The definition states that if $[K \Rightarrow K']_L$, i.e. the translation of $[K \Rightarrow K']$ to some knowledge representation formalism L, follows from WK_L , i.e. the world knowledge specified in this language L, with regard to a notion of validity (\models_L) defined on L and in addition K_1 implies K'_1 with regard to $[K \Rightarrow K']$, i.e. $K_1 \Rightarrow_{[K \Rightarrow K']} K'_1$, and furthermore K_α is *m-suitable* to K'_1 , then K_α will be considered as a bridging reference to K_1 such that the inferred DRS K'_1 is incorporated into the main DRS thus making the relation between K_1 and K_α explicit.

The benefit of this modification is that we have abstracted from Bos et al.'s qualia structure thus allowing any form of knowledge representation formalism L as well as a corresponding inference mechanism to be integrated into the anaphora resolution process. In fact, for doing so, only the translation from DRT to L as well as the notion of validity (\models_L) has to be defined. For instance, if first order logic (FOL) is the language L and consequently \models_{FOL} denotes the validity of first order logic formulas, Blackburn et al.'s idea of proving the validity of a first order logic formula by running simultaneously different theorem provers ([5]) could be exploited within bridging reference resolution.

Let's return to our example (17) and analyze the second sentence with the definite description 'THE LBP/LPS COMPLEX'. The following α -marked DRS K" will be assumed for the definite description:

$K'' : \alpha :$	<table border="1"> <tr> <td>s'', c'', p_1'', p_2''</td> </tr> <tr> <td> $state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$ </td> </tr> </table>	s'', c'', p_1'', p_2''	$state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$
s'', c'', p_1'', p_2''			
$state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$			

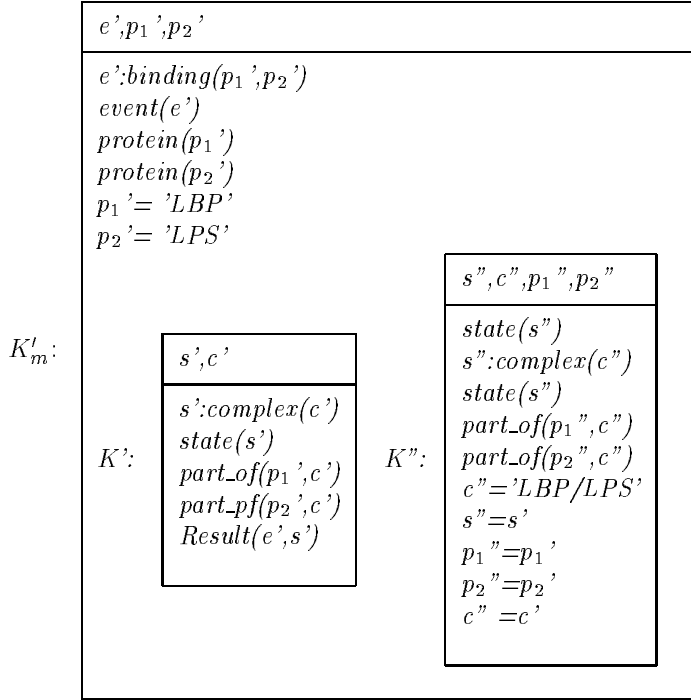
First of all this DRS has to be merged with the DRS for the first sentence thus yielding:

$K_m :$	<table border="1"> <tr> <td>e', p_1', p_2'</td> </tr> <tr> <td> $e' : binding(p_1', p_2')$ $event(e')$ $protein(p_1')$ $protein(p_2')$ $p_1' = 'LBP'$ $p_2' = 'LPS'$ </td> </tr> <tr> <td style="padding-left: 20px;">$K'' : \alpha :$</td> <td> <table border="1"> <tr> <td>s'', c'', p_1'', p_2''</td> </tr> <tr> <td> $state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$ </td> </tr> </table> </td> </tr> </table>	e', p_1', p_2'	$e' : binding(p_1', p_2')$ $event(e')$ $protein(p_1')$ $protein(p_2')$ $p_1' = 'LBP'$ $p_2' = 'LPS'$	$K'' : \alpha :$	<table border="1"> <tr> <td>s'', c'', p_1'', p_2''</td> </tr> <tr> <td> $state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$ </td> </tr> </table>	s'', c'', p_1'', p_2''	$state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$
e', p_1', p_2'							
$e' : binding(p_1', p_2')$ $event(e')$ $protein(p_1')$ $protein(p_2')$ $p_1' = 'LBP'$ $p_2' = 'LPS'$							
$K'' : \alpha :$	<table border="1"> <tr> <td>s'', c'', p_1'', p_2''</td> </tr> <tr> <td> $state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$ </td> </tr> </table>	s'', c'', p_1'', p_2''	$state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$				
s'', c'', p_1'', p_2''							
$state(s'')$ $s'' : complex(c'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$							

Now it is the case that K'' is m -suitable to the DRS K' which has been proved to follow from the DRS K (see above), so that according to definition (14) K'' will be interpreted as a bridging reference to K and thus resolved to the DRS:

$K'' :$	<table border="1"> <tr> <td>s'', c'', p_1'', p_2''</td> </tr> <tr> <td> $state(s'')$ $s'' : complex(c'')$ $state(s'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$ $s'' = s'$ $p_1'' = p_1'$ $p_2'' = p_2'$ $c'' = c'$ </td> </tr> </table>	s'', c'', p_1'', p_2''	$state(s'')$ $s'' : complex(c'')$ $state(s'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$ $s'' = s'$ $p_1'' = p_1'$ $p_2'' = p_2'$ $c'' = c'$
s'', c'', p_1'', p_2''			
$state(s'')$ $s'' : complex(c'')$ $state(s'')$ $part_of(p_1'', c'')$ $part_of(p_2'', c'')$ $c'' = 'LBP/LPS'$ $s'' = s'$ $p_1'' = p_1'$ $p_2'' = p_2'$ $c'' = c'$			

Furthermore, the DRS K' will be accommodated into the main DRS such that finally we yield:



So finally the resultative relation between the state s'' of being complexed and the binding event e' has been made explicit through accommodation of the inferred DRS K' and accommodation of the binding relations yielded by the resolution of K'' , i.e. $Result(e', s'')$ follows from $Result(e', s')$ and $s'' = s'$.

4.2 Reasoning within an Ontology

The goal of the work presented here is to make use of an ontology as backbone of the inference mechanism driving the bridging resolution process. For this purpose it has to be formally defined what an ontology is. Furthermore, a suitable declarative language has to be found to represent it and a notion of validity has to be introduced by defining an inference method on this language. An ontology is a specification of a conceptualization ([19]). A conceptualization can be understood as an abstract representation of the world or domain we want to model for a certain purpose. The author's idea of an ontology is completely in line with Gruber's ([19]):

Formally, an ontology is the statement of a logical theory.
Ontologies are often equated with taxonomic hierarchies of classes, class definitions and the subsumption relation, but ontologies need

not be limited to these forms. Ontologies are also not limited to *conservative definitions*, that is, definitions in the traditional sense that introduce only terminology and do not add any knowledge about the world. To specify a conceptualization one needs to state axioms that do constrain the possible interpretations for the defined term.

According to Gruber's ideas, an ontology will be understood as a specification of a conceptualization of a domain through:

1. a taxonomic hierarchy of concepts
2. partial definitions of concepts (these are axiomatic definitions specifying the necessary conditions of a concept)

Furthermore, following Gruber's advice: "When a definition can be stated in logical axioms, it should be", first order logic will be used as declarative representation language. Formally, an ontology will be defined as follows:

Definition 15 (Ontology)

An ontology O is a triple (C, T, D) where C is a set of predicates representing concepts, T is a set of first order logic formulas representing taxonomic relations between the concepts in C and D is a set of first order logic formulas defining the concepts in C .

Let's illustrate the above definition with two examples. The fact that a chandelier is some sort of lamp is such a taxonomic relation in the above sense and can be expressed through the following formula:

$$\forall x (chandelier(x) \rightarrow lamp(x)) \tag{18}$$

On the other hand the fact that rooms have lamps can be seen as a partial definition of a room as it is in its nature to have lamps. This partial definition can be expressed as follows:

$$\forall x (room(x) \rightarrow \exists y (lamp(y) \wedge part_of(y, x))) \tag{19}$$

And here's the tiny ontology we have built containing world knowledge about rooms¹⁰:

- $O_{room} = (C_{room}, T_{room}, D_{room})$
- $C_{room} = \{\lambda x.room(x), \lambda x.chandelier(x)\}$

¹⁰Certainly, the semantics of the part-of relation should also be defined in form of logical axioms stating its anti-symmetry as well as its transitivity. This can be done within a separate set of inference rules as defined in ([40]). However, the author will not go any further into this aspect within the work presented here.

- $T_{room} = \{\forall x (chandelier(x) \rightarrow lamp(x))\}$
- $D_{room} = \{\forall x (room(x) \rightarrow \exists y (lamp(y) \wedge part_of(y, x)))\}$

From a logical point of view there is no difference between taxonomic relations and axiomatic definitions of concepts as they will both be universally quantified implications, but there is a reason for treating them separately which will be clear when giving an ontology based definition of the bridging resolution algorithm of Bos et al.

Having defined what an ontology is, it should be clarified how an inference mechanism can be defined on such an ontology. In order to exploit the taxonomic relations T defined within the ontology in the bridging resolution process, a notion of specialization between DRSs will be introduced. In this sense a DRS K_1 is a specialization of a DRS K_2 if K_1 represents an ontological concept which is more special than the one represented by K_2 :

Definition 16 (Specialization)

A DRS K_1 is a specialization of a DRS K_2 with regard to an ontology $O=(C,T,D)$ ($K_1 <_O K_2$) iff \exists DRS $[K'_1 \Rightarrow K'_2]$ and $[K'_1 \Rightarrow K'_2]_{PL_1} \in T$ and $K_1 \Rightarrow_{[K'_1 \Rightarrow K'_2]} K_2$, where $[K]_{PL_1}$ is the translation of the DRS K to predicate logic as defined in [26].

According to the above definition the DRS

y
$chandelier(y)$

 for example is a

specialization of the DRS

y
$lamp(y)$

 with regard to the ontology O_{room} .

In the following, \leq_O^* will denote the reflexive and transitive closure of $<_O$. Now it can be defined what it means for a DRS K' to follow logically from K within an ontology O . Intuitively, a DRS K' will follow from K with regard to the ontology if there is a DRS K'' such that K is more special than K'' and K' follows from K'' by a conceptual definition in D :

Definition 17 (Implication within an ontology)

A DRS K implies a DRS K' within an ontology $O=(C,T,D)$ (Notation: $K \Rightarrow_O K'$) iff \exists DRS $[K_1 \Rightarrow K_2]$ such that $[K_1 \Rightarrow K_2]_{PL_1} \in D$ and $K \Rightarrow_{[K_1 \Rightarrow K_2]} K'$ or \exists DRS K'' such that $K \leq_O^* K''$ and $K'' \Rightarrow_O K'$.

The inference mechanism defined in this way boils down to substituting the restrictor of the complex condition $[K_1 \Rightarrow K_2]$ determining the forward-chaining by a DRS with a smaller denotation, i.e. more special with regard to the \leq_O^* -relation. It is important to mention that this is only possible if the complex

condition $\boxed{K_1 \implies K_2}$ has a positive polarity and the restrictor consequently a negative polarity as defined in [57]. Within the work presented in this thesis it will in fact be assumed that all the complex conditions within the ontology O have a positive polarity such that the restrictor can always be replaced with a DRS representing a concept which is more special, i.e. with a smaller denotation.

Regarding the ontology O_{room} , it is for example the case that the DRS

y
$room(y)$

implies the DRS

z
$lamp(z)$ $part-of(z, y)$

It thus seems that we have all the ingredients for a successful resolution of example (3) as it can be inferred with regard to our ontology that rooms have lamps as parts and that a chandelier is some sort of lamp. The remaining question is how the ontology-based inference mechanism defined here can be integrated into the bridging reference resolution process.

4.3 Integrating the Ontology

The interpretation of the definite description 'the chandelier' in example (3) as being part of the room mentioned in the first sentence is only possible if:

1. the ontology allows us to infer that rooms have lamps
2. the ontology's taxonomy allows us to infer that chandeliers are lamps
3. and the anaphora resolution algorithm allows us to relate the inferred lamp in the first sentence to the 'taxonomically' inferred lamp in the second one

In the last section it has been shown that the first two conditions can be fulfilled by defining an appropriate inference mechanism on the basis of a certain ontology. The third condition reveals that bridging reference resolution has to be understood and thus also modeled as an interplay between exploiting knowledge in form of axiomatic definitions of concepts and taxonomic reasoning. And this is exactly the reason why an ontology seems so suitable for the purpose. As mentioned before, both types of reasoning are basically logical inferences. The reason why in the approach described here they are treated separately is in fact a question of linguistic nature. To illustrate this, let's for convenience recall example (3):

(20) I walked into the room. <The chandelier(s)> sparked brightly.

Certainly, the lamp inferred from the room as well as the fact that it is part of it has to be accommodated such that we can refer to it, but the (taxonomically inferred) lamp in the second sentence does not need to be accommodated as

this would lead to an unnecessary redundancy because *the chandelier* already represents a specific type of lamp.

In fact, following Van der Sandt it would even be wrong to accommodate it as it would not represent any new information. In Van der Sandt's view, accommodation is subject to various conditions. First, it should not introduce free variables. Second, it should be consistent, i.e. not introduce any contradictions. Third, it should also be informative in the sense that the accommodated material is not entailed by the preceding context. Now it should be clear why taxonomic relations and axiomatic definitions of concepts have been treated separately. Axiomatic definitions of concepts introduce knowledge which should be accommodated, while taxonomic generalizations do not introduce new information and thus should not be accommodated.

The example (3) corresponds to a pattern that could be paraphrased as 'referring to an antecedent which can be inferred from the previous context'. In fact, the classes of bridging references *Set membership*, *Necessary parts*, *Probable parts*, *Inducible parts*, *Necessary roles*, *Optional roles*, *Reasons*, *Causes*, *Consequences* and *Concurrences* defined by Clark follow the above pattern. The relation between the referring expression and the inferred antecedent is then *set membership*, *part-of* or *role-of* for referring entities and *elaboration*, *explanation*, *result* or *parallel* for events or states. However, the above pattern is not the only possible one. The pattern of 'referring back to a previous expression (in a more general way)' is also quite common. The corresponding class within Clark's taxonomy of bridging references is clearly the one he calls *Identity*. Here follow some examples for this pattern from the SWISS-PROT corpus:

- (21) BINDS TO THE METAL-REGULATING-ELEMENT (MRE) OF METALLOTHIONEIN IA GENE PROMOTER. <BINDING> IS ZINC-DEPENDENT.
- (22) INCREASES THE FORMATION OF RIBOSOMAL TERMINATION COMPLEXES AND STIMULATES ACTIVITIES OF RF-1 AND RF-2. [...] <THE STIMULATION OF RF- 1 AND RF-2> IS SIGNIFICANTLY REDUCED BY GTP AND GDP, BUT NOT BY GMP.
- (23) BINDS TO THE CYTOPLASMIC DOMAIN OF THE CELL-CELL ADHESION MOLECULE E-CADHERIN, AND PERHAPS TO OTHER (MEMBRANE) PROTEINS. <THE ASSOCIATION OF CATENINS TO CADHERINS> PRODUCES A COMPLEX WHICH IS LINKED TO THE ACTIN FILAMENT NETWORK, AND WHICH SEEMS TO BE OF PRIMARY IMPORTANCE FOR CADHERINS CELL-ADHESION PROPERTIES.

The following definition of the *bridging*-operation captures both of the patterns paraphrased above in a formal way making use of the definitions of specialization and implication presented in section 4.2:

Definition 18 (Bridging)

$$\begin{aligned}
\text{BRIDGE}(K_\alpha, K_m) = \{ & K'_m \mid K_2 \leq K_1 \leq K_m \wedge \alpha : K_\alpha \in C(K_2) \wedge \\
& K_1 \implies_O K'_1 \wedge K_\alpha \leq_O^* K'_1 \wedge \\
& K'_\alpha \text{ } m\text{-suitable to } K'_1 \wedge \\
& U(K_3) = U(K_2) \cup U(K_\alpha) \wedge \\
& C(K_3) = C(K_2) - \alpha : K_\alpha \cup C(K_\alpha) \cup \{x = y \mid m(x) = y\} \wedge \\
& U(K_4) = U(K_1) \cup U(K'_1) \wedge \\
& C(K_4) = C(K_1) \cup C(K'_1) \wedge \\
& K'_m = K_m[K_2/K_3][K_1/K_4] \} \\
\cup \{ & K'_m \mid K_2 \leq K_1 \leq K_m \wedge \alpha : K_\alpha \in C(K_2) \wedge \\
& K_1 \leq_O^* K'_1 \wedge K_\alpha \text{ } m\text{-suitable to } K'_1 \wedge \\
& U(K_3) = U(K_2) \cup U(K_\alpha) \wedge \\
& C(K_3) = C(K_2) - \alpha : K_\alpha \cup C(K_\alpha) \cup \{x = y \mid m(x) = y\} \wedge \\
& K'_m = K_m[K_2/K_3] \}
\end{aligned}$$

The first part of the above definition captures the first of the two patterns introduced above. It states that in case a DRS K'_1 follows logically from a DRS K_1 with regard to the ontology O and K'_α is m -suitable to the DRS K'_1 , where K_α is a specialization of K'_α , K_α will be considered as a bridging reference to K_1 and the 'inferred' DRS K'_1 will be accommodated so that the relation between K_α and K_1 becomes explicit.

The same explanation holds for the second part of the definition which corresponds to the second pattern. The only difference is that K'_1 is merely an ontological generalization of K_1 such that it will not be accommodated in line with Van der Sandt's view.

4.4 Determinacy and Reasoning Complexity

The author agrees with Clark that in natural language discourse bridges are determinate in the sense that readers/listeners always build the shortest path with regard to their memory and world knowledge to yield a specific resolution. Thus in Clark's view determinacy in bridging resolution boils down to reducing the effort to establish a certain bridge between a referring expression and its antecedent. A formal model of bridging should definitely account for this determinacy. Within the ontology driven approach presented here, bridges will be made determinate by choosing the most recent antecedent as possible and by minimizing reasoning complexity. The complexity of the reasoning within an ontology as defined within this work is basically determined by the number of generalizations needed to infer a suitable antecedent. In fact, it follows from the bridging definition (18) that in the worst case¹¹ the reasoning complexity will amount to $|A|^2 + 1$ applications of the Generalized Modus Ponens (GMP). Thus the complexity of the inference mechanism presented is $O(|A|^2)$, i.e. quadratic in the number of specialization/generalization relations $|A|$ of the ontology. So, the bridging resolution will be made determinate by choosing the

¹¹This is the one in which all the concepts are linearly ordered with regard to the $<_O$ -relation thus forming a degenerated 'ontological tree' consisting of only one 'branch'.

most recent antecedent and minimizing the number of generalizations and thus also applications of the GMP.

5 The Ontology

5.1 Pathways: the General Picture

As already mentioned in section 1, the aim of the work presented here has been to investigate and discover the principles underlying and driving the resolution of bridging references within information extraction systems. For this purpose, a domain-independent and ontology-driven approach as described in section 4 has been developed. However, the verification of this approach as well as of the underlying principles discovered can only be conducted with regard to a specific domain. In order to apply the approach to the domain of molecular biology, it has been crucial to built up domain knowledge concerning the function of proteins as well as their possible interactions. Within this respect the modeling of *pathways* has turned out to be a key issue. A *pathway* can be defined as the ordered and synchronized activity of different proteins which as a whole has a specific function ([8]). The activities of single proteins reach from simple bindings over certain chemical modifications such as phosphorylation, hydrolysis, etc. through to complex interactions such as transport events from one cellular location to another. Thus from a narrow point of view the function of a single protein is to accomplish a certain step within some pathway. But in a wider sense the function of a protein can also be specified with regard to the consequences its activity has within a certain pathway. If a protein for example activates some other protein by phosphorylating it and the latter then acts as a transcription factor stimulating transcription, we can definitely assert that the former protein (in a wider sense) stimulates transcription. So when talking about interactions of proteins we have to bear in mind that the function of this interaction can be seen from a 'narrow perspective' as well as from an 'overall perspective' by considering the pathway as a whole and also the sum of all the pathways which determine the behavior of the whole organism. Within this section a conceptual model of the interactions between proteins as well as the relations between them will be presented. This model has to be regarded as a very limited one which nevertheless captures some of the interactions which make up a typical pathway. The insights on which this model is grounded have been obtained by the author basically through:

- previous work on the conceptual modeling of the role of RNA polymerase within gene transcription ([36])
- the participation in a project with the aim of developing an ontology of biochemical compounds, reactions and pathways ([58])
- intense corpus studies (SWISS-PROT, JBC)
- personal communication with several biologists

Figure (1) shows part of the developed model in form of a graph. The nodes of the graph represent events which have turned out to be crucial for the modeling of pathways, while the edges between them represent conceptual relations

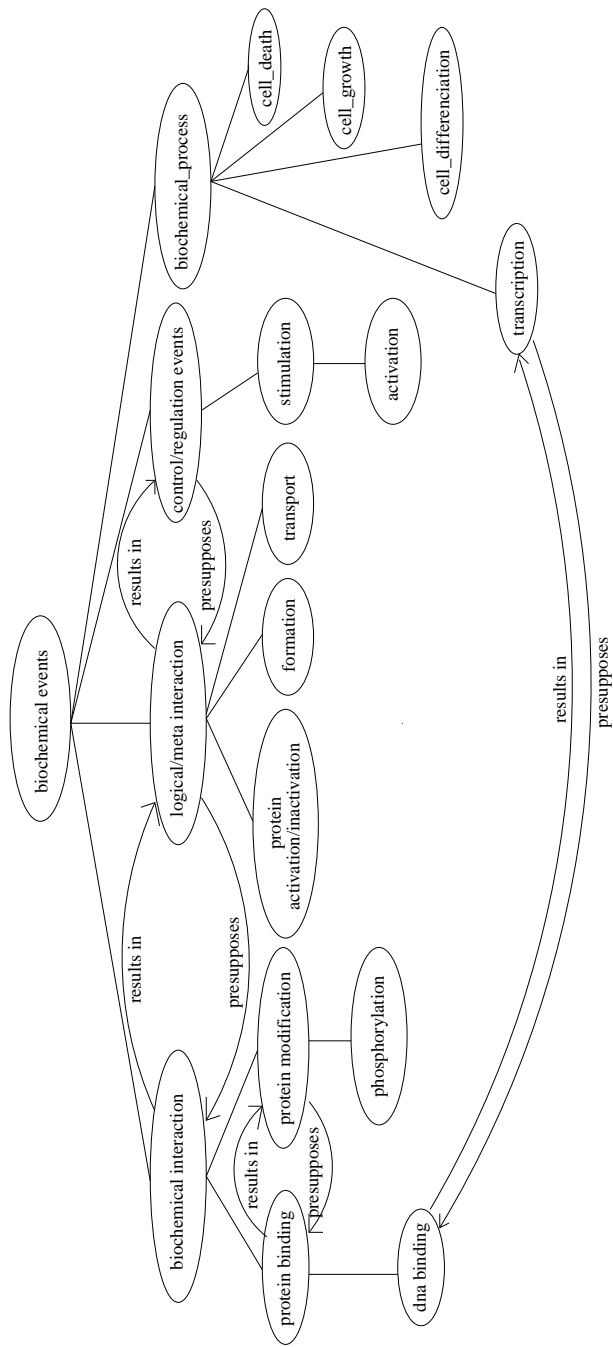


Figure 1: Model of protein interactions

between these events. The edges without an arrow represent taxonomic or generalization/specialization relations (also widely called *is-a* relations). The concepts are ordered top-down from most general to most specific. The other edges are marked with the relations *results in* and *presupposes*. No formal semantics has been defined for these relations because an intuitive explanation should suffice for the purposes here. The intuitive meaning of an edge labeled with the relation *results in* is that within a pathway an event of the first type normally leads to an event of the second type. Similarly, the meaning of an edge labeled with the relation *presupposes* is intended to mean that the existence of an event of the first type presupposes the existence of one of the second type. The aim of this graph is to visualize the model or general picture underlying this work. It may be argued that it is a very limited and simple one, but maybe exactly because of this also transparent enough to illustrate the application of the bridging reference resolution approach to the domain in question. The picture drawn is essentially one in which proteins interact in some way to control or regulate certain biochemical processes such as cell death, cell growth, differentiation or transcription of genes (yielding proteins which in turn interact in some way). In particular, the biochemical interaction of two proteins can lead to the catalytic activation or inactivation of one of them or to the formation of a complex or a phosphate bond and this in turn can stimulate or inhibit - i.e. control positively or negatively - other biochemical processes. A binding of two proteins, which is a specific type of interaction, can as a consequence lead to the modification (phosphorylation, hydrolysis, methylation) of one of them while such a modification always presupposes a binding between the proteins in question. In a similar way a binding can also lead to a transport of the bound protein and a protein which binds to a DNA molecule can act as a transcription factor thus leading to the transcription of a specific gene. In both cases a corresponding binding has to be considered as a necessary condition.

To conclude this section, the model outlined above will be applied to the well-known JAK/STAT pathway ([30]) representing the key interactions in terms of the events and relations introduced above. As shown in figure 2, within this pathway the binding of the ligand to the Cytokine receptors leads to the autophosphorylation and activation of JAK and in turn to the phosphorylation of the Cytokine receptors which then phosphorylate STAT. As a result STAT then forms homodimers and is translocated to the nucleus where it binds to DNA and activates the transcription of target genes.

It is important to insist on the fact that the aim of the picture described above is not to explain the general mechanism of a typical pathway in an exhaustive and detailed manner. The aim has been to 'draw' a rather 'impressionistic' picture of some crucial protein interactions and biochemical events as well as relations between them which typically take place in pathways. In contrast, the next section will create a more 'realistic' picture by picking out specific classes of events and defining their semantics. The sum of these two pictures will then hopefully yield a model of our domain in form of an ontology detailed and broad enough for the bridging resolution process to be successful.

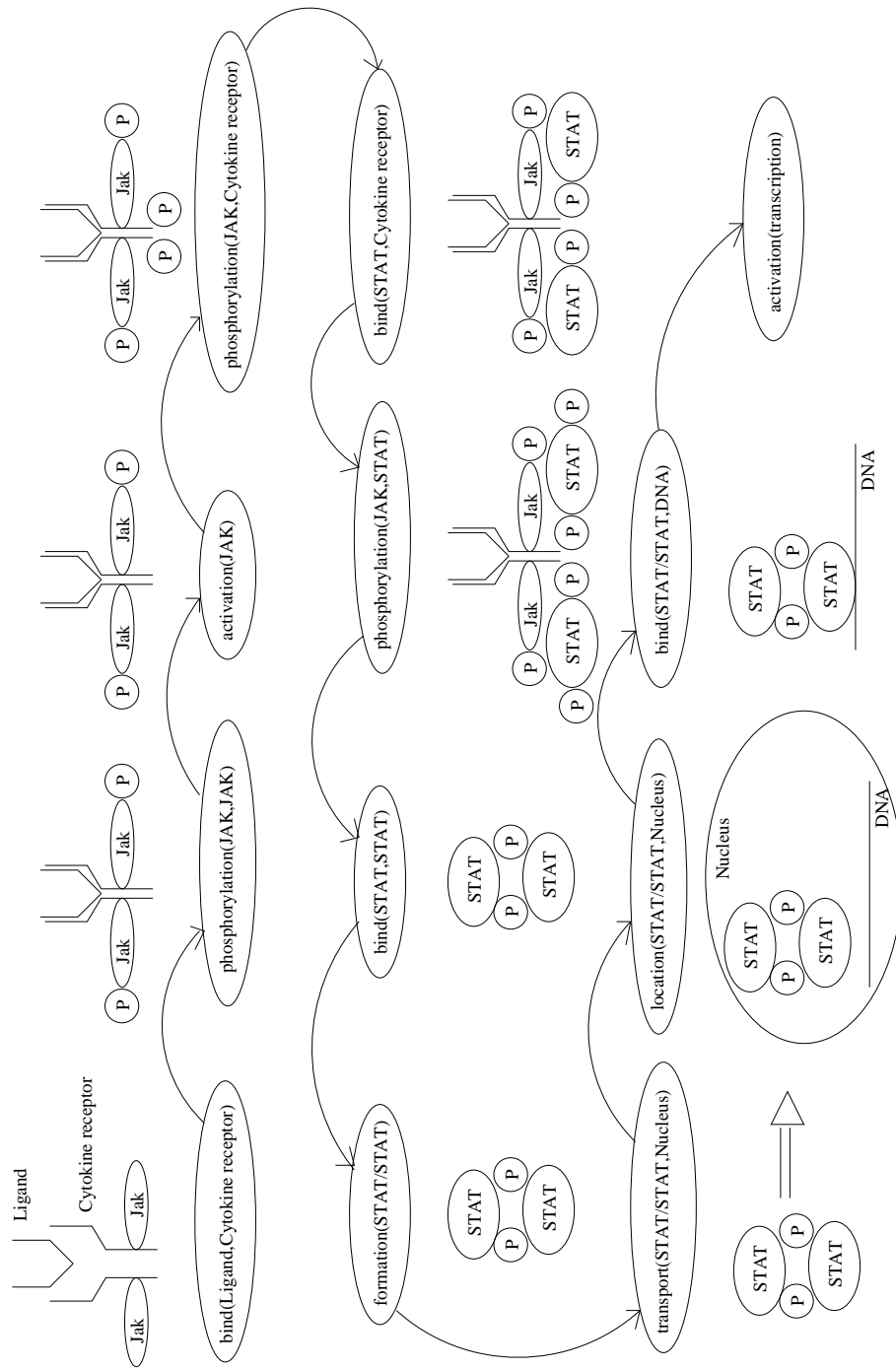


Figure 2: The JAK/STAT pathway in 12 steps

5.2 Event or not event?

Events will be understood within the work presented here in line with ([56]) as representing a change. A change will be regarded as a transition from a state s in which some condition p holds to a state s' where p does not hold anymore, i.e. $\neg p$ holds.

The author is aware of the fact that it may be argued that not all events involve changes, but will not discuss this possibility further as it would probably lead to a philosophical terrain the author would merely stumble along.

The view of events underlying this work is that they represent entities as argued by Davidson ([14]) and furthermore involve an observable or measurable change to the world they exist in. In fact, the author does not know how to justify the existence of an event which is not observable at all. States, as opposed to events, will be considered as not involving any change but implying the continuation of a certain condition over a time interval ([26]). Having defined what an event is and how it differs from a state, the importance of this distinction should be discussed. From a linguistic point of view, such a distinction is important for discourse analysis and in particular for the computation of discourse relations ([34], [35]). In fact, the knowledge whether something is an event or a state poses constraints on the rhetorical relation to a potential antecedent. For example, an event can never be understood as standing in a *background* relation with an antecedent representing some other event because *background* as defined by Asher et. Lascarides ([34],[35]) is a relation between a state and an event which temporally overlap.

On the other hand, identifying something as an event as well as its pre- and postconditions allows to group events semantically and conceptually into classes. Furthermore, if a taxonomic hierarchy for these conditions is available, we will consequently also yield a taxonomic hierarchy of events by identifying the events with more specific pre- and postconditions as subclasses of events with less specific ones. Such a hierarchical classification of events has a direct application in ontology design. This possibility will be discussed in detail in section 5.3.2. From a logical and deductive point of view, the fact that a certain event e has taken place allows us to infer that before e the corresponding condition p and after e $\neg p$ were true, while a state does not allow such an inference. This has direct applications in database update ([31]).

5.2.1 Linguistic Realization of Events

Events are linguistically expressed by active or passive verbal forms as well as by nominal phrases. The following expressions for example are linguistic realizations of one and the same event:

- JAK phosphorylates STAT
- the phosphorylation of STAT by JAK
- STAT is phosphorylated by JAK

Thus events have to be seen as meta-linguistic entities so that their representation should be independent of the various linguistic realizations by which it can be expressed.

The task of mapping a verb to a corresponding event is not always a straightforward one. This is due to several reasons. First, the verb may in fact represent a state rather than an event. Second, it may turn out to be rather difficult to specify the pre- and postconditions that justify it as such. Third, some verbs may have a generic as well as an 'event' reading. All these cases apply for example to the verb *control*. The sentence 'JAK controls transcription.' has a generic reading denoting the property of JAK to control transcription (thus representing a state) as well as an 'event' reading in which a *control* event actually takes place. In the work presented here, the generic reading of verbs such as 'control' will be ignored. But even if we merely consider the event reading of certain verbs, it can turn out to be quite difficult to specify the pre- and postconditions of the corresponding event. In such cases the problem will be evaded by the use of second order logic to state that there exists some property or condition which is changed by the event.

5.2.2 A Semantic Classification of Events

An intense corpus study allowed to identify 11 different semantic classes of events together with their corresponding pre- and postconditions. The assignment of a certain event to a specific class is to a certain extent arbitrary in the sense that the indicated pre- and postconditions are far from being complete and in principle do not exclude each other so that the classes are not necessarily disjoint.

Furthermore, many different events can be linguistically realized by one and the same verb (through different subcategorization structures). A verb like *control* has in fact different meanings and therefore represents also different events $control_1, \dots, control_n$. The classification proposed here is neither complete with regard to all the events represented by a certain verb nor with regard to the specification of the pre- and postconditions. Anyway, the aim of this classification is certainly not completeness. Its aim is merely to justify the event reading of a certain verb and to allow a rough semantic classification of the verbs within the considered domain. In the following, the different classes will be presented together with some examples from the corpus.

Control/Regulation From an 'overall perspective' it is certainly correct to assert that proteins usually control or regulate biochemical processes in the sense that they affect them by stimulating, activating or inhibiting them. However, it seems quite difficult to specify the pre- and postconditions of such a *control/regulation*-event. A control/regulation can actually affect the speed as well as the concentration of the products or other properties of a certain reaction or biochemical process. Thus a *control/regulation*-event can be intuitively formalized as changing the value of some property of a certain biochemical pro-

cess:

$$\begin{aligned}
& \forall e, bp, p_1 ((e : control(p_1, bp) \wedge protein(p_1) \wedge biochemical_process(bp)) \Leftrightarrow \\
& \exists P, s, s' (measurable_property(P) \wedge \\
& s : value(P(bp)) = v_1 \wedge s \supset e \wedge \\
& s' : value(P(bp)) = v_2 \wedge v_2 \neq v_1 \wedge Result(e, s'))
\end{aligned} \tag{24}$$

In the above formula $Result(e, s')$ is the rhetorical relation defined by Lascarides et al. ([34]) and its meaning is that the event e causes the event or state s' . The axiomatic definition of $Result$ used within this master's thesis will be the following:¹²

$$\begin{aligned}
& \forall e, e' ((Result(e, e') \wedge event(e) \wedge eventuality(e')) \rightarrow \\
& (cause(e, e') \wedge e \supset e'))
\end{aligned} \tag{25}$$

On the other hand $Explanation(e, e')$ can be understood as the inverse relation to $Result$. Thus its axiomatic definition is as follows:

$$\begin{aligned}
& \forall e, e' ((Explanation(e, e') \wedge event(e) \wedge eventuality(e')) \rightarrow \\
& (cause(e', e) \wedge e' \supset e))
\end{aligned} \tag{26}$$

The following table contains verbs denoting *control/regulation*-events which change some property of a biochemical process. If a specific property changed by the event can be indicated, it will be given in the second column. Furthermore, for some events it is possible to specify a relation between the values v_1 and v_2 of the changed property. If this is possible, this relation will be given in the third column of the table:

Control/Regulation			
event	property	relation	example
accelerate	speed(bp)	$v_2 > v_1$	<u>ACCELERATES</u> PROGRAMED CELL DEATH BY BINDING TO, AND ANTAGONIZING THE APOPTOSIS REPRESSOR BCL-2 OR ITS ADENOVIRUS HOMOLOG E1B 19K PROTEIN.
antagonize	effect(bp)	$v_2 < v_1$	HESX1 CAN <u>ANTAGONIZE</u> PROP1 ACTIVATION.
augment		$v_2 > v_1$	ABLE TO <u>AUGMENT</u> THE PROLIFERATION OF BOTH MYELOID AND LYMPHOID HEMATOPOIETIC PROGENITORS IN BONE MARROW CULTURE.
catalyze			<u>CATALYZES</u> THE ANAEROBIC DEHYDRATION OF L-CARNITINE (R-3-HYDROXY-4-AMINO BUTYRATE) TO CROTONOBETAINE.

¹²The question of how to define the notion of causality and thus the semantics of the *cause* relation are not discussed within this master thesis. The reader is referred to [38].

Control/Regulation			
event	property	relation	example
control			THIS PROTEIN <u>CONTROLS</u> THE EXPRESSION OF AT LEAST SIX GENES THAT ARE INVOLVED IN THE TRANSPORT AND CATABOLISM OF L-ARABINOSE.
coordinate			MAY <u>COORDINATE</u> MEMBRANE TRANSPORT WITH THE FUNCTION OF THE CYTOSKELETON.
cooperate			<u>COOPERATES</u> WITH MYOD IN THE ACTIVATION OF CARDIAC ACTIN GENE EXPRESSION.
decrease		$v_2 < v_1$	JH ESTERASE PLAYS A CRUCIAL ROLE IN THE <u>DECREASE</u> OF JH ACTIVITY IN LEPIDOPTERAN INSECTS, BY HYDROLYZING THE METHYL ESTER OF JH.
drive			RECRUITS MITOCHONDRIAL HSP70 TO <u>DRIVE</u> PROTEIN TRANSLOCATION INTO THE MATRIX USING ATP AS AN ENERGY SOURCE.
enhance		$v_2 > v_1$	COULD <u>ENHANCE</u> THE INCORPORATION OF NICKEL TO THE HYDROGENASE.
increase		$v_2 > v_1$	<u>INCREASE</u> GTP HYDROLYSIS INDUCED BY THE RAN GTPASE ACTIVATING PROTEIN RANGAP1.
influence			ROB BINDING MAY <u>INFLUENCE</u> THE FORMATION OF THE NUCLEOPROTEIN STRUCTURE, [...]
modulate			IT MAY <u>MODULATE</u> THE ASSOCIATION OF TROPOMYOSIN WITH THE SPECTRIN-ACTIN COMPLEX IN THE ERYTHROCYTE MEMBRANE SKELETON, [...]
(auto-/co-)regulate			<u>REGULATES</u> GENE EXPRESSION THROUGH ACTIVATION OF NF κ B.
stimulate		$v_2 > v_1$	BINDS TO PLATELET GPIB/IX RECEPTOR SYSTEM AND STIMULATES AGGLUTINATION.

Activation or inhibition events can be seen as a special case of a *control/regulation*-event in the sense that $v_1 = 0$ in the case of an activation and $v_2 = 0$ in the case of an inhibition. The following tables contain verbs respectively denoting activation and inhibition events:

activation($v_1 = 0$)	
event	example
activate	<u>ACTIVATES</u> THE EXPRESSION OF DRUG METABOLIZING ENZYMES GENES (SUCH AS THE CYP1A1 GENE).
allow	IT IS THOUGHT TO <u>ALLOW</u> RNA POLYMERASE READ THROUGH A RHO- INDEPENDENT TRANSCRIPTION TERMINATOR BETWEEN THE AMIE PROMOTER AND GENE.
cause	ELICIT LEAF NECROSIS AND <u>CAUSE</u> THE ACCUMULATION OF PATHOGENESIS-RELATED PROTEINS.
elicit	EMAP II <u>ELICITS</u> A PHLOGOGENIC RESPONSE AND, POTENTIALLY, AUGMENTS THE EFFECTS OF THE OTHER TUMOR-DERIVED CYTOKINES.
facilitate	<u>FACILITATES</u> RIBOSOME BINDING BY INDUCING THE UNWINDING OF THE MRNAS SECONDARY STRUCTURES.
induce	CAN <u>INDUCE</u> B-CELL PROLIFERATION, [...]
initiate	<u>INITIATES</u> LUTEOLYSIS IN THE CORPUS LUTEUM.
lead	THE REACTION CATALYZED BY TOPOISOMERASES <u>LEADS</u> TO THE CONVERSION OF ONE TOPOLOGICAL ISOMER OF DNA TO ANOTHER.
mediate	<u>MEDIATES</u> T5 BINDING TO ITS HOST RECEPTOR, THE E.COLI FHUA PROTEIN.
promote	IT IS A SINGLE- STRANDED DNA BINDING PROTEIN THAT CAN <u>PROMOTE</u> RENATURATION OF DNA.
respond	MAY PLAY AN IMPORTANT ROLE IN THE CAPACITY OF CELLS TO <u>RESPOND</u> TO AND COPE WITH CHANGES IN THEIR ENVIRONMENT.
result	DURING VIRAL ASSEMBLY, THE PROTEINS FORM MEMBRANE ASSOCIATIONS AND SELF-ASSOCIATIONS THAT ULTIMATELY <u>RESULT</u> IN BUDDING OF AN IMMATURE VIRION FROM THE INFECTED CELL.
transactivate	INVOLVED IN <u>TRANSACTIVATING</u> ANAEROBIC EXPRESSION OF THE PHOTOSYNTHETIC APPARATUS.
trigger	ISOFORM IIB2 DOES NOT <u>TRIGGER</u> PHAGOCYTOSIS.

inhibition($v_2 = 0$)	
event	example
block	IT <u>BLOCKS</u> ELONGATION OF THE TRANSCRIPTION OF HOST DNA AND OF OTHER PHAGES.
hinder	REPRESSION OCCURS BY STERICALLY <u>HINDERING</u> THE ASSEMBLY OF THE TRANSCRIPTION INITIATION COMPLEX.
inhibit	ANTIMICROBIAL PEPTIDES THAT <u>INHIBIT</u> THE GROWTH OF NUMEROUS SPECIES OF BACTERIA AND FUNGI AND INDUCE OSMOTIC LYSIS OF PROTOZOA.
repress	BINDS TO THE P2 PROMOTER REGION OF THE C-MYC GENE AND IS THOUGHT TO <u>REPRESS</u> TRANSCRIPTION.
suppress	IT ALSO <u>SUPPRESSES</u> A VARIETY OF MITOCHONDRIAL INTRON MUTATIONS.
terminate	<u>PUTATIVELY TERMINATES</u> POLYKETIDE BIOSYNTHESIS BY TRANSFER OF POLYKETIDE ACYL CHAIN TO THE SECONDARY AMINE ON THE 3- AHBA STARTER UNIT.

Biochemical Interactions Chemical substances interact with each other in various ways. These interactions normally result in the change of some biochemical property of some of the implicated substances:

$$\begin{aligned}
& \forall e, s_1, s_2 ((e : \text{biochemical_interaction}(s_1, s_2) \wedge \\
& \text{chemical_substance}(s_1) \wedge \text{chemical_substance}(s_2)) \Leftrightarrow \\
& \exists P, s, s' (\text{biochemical_property}(P) \wedge (s : \neg P(s_1) \wedge s' : P(s_1) \vee \\
& s : \neg P(s_2) \wedge s' : P(s_2)) \wedge s \supset \subset e \wedge \text{Result}(e, s'))) \quad (27)
\end{aligned}$$

The following table contains verbs representing biochemical interactions. The column in the middle indicates the corresponding biochemical property P which is changed:

Biochemical Interactions		
event	P	example
cleave	$\neg \text{cleaved}(s)$	WHEN A PROTEINASE <u>CLEAVES</u> THE BAIT REGION, A CONFORMATIONAL CHANGE IS INDUCED IN THE PROTEIN WHICH TRAPS THE PROTEINASE.
depolarize	$\neg \text{depolarized}(s)$	THIS CLASS OF TRANSMEMBRANE TOXINS <u>DEPOLARIZE</u> THE CYTOPLASMIC MEMBRANE, [...]
hydrolyze	$\neg \text{hydrolyzed}(s)$	THIS ENZYME IS A THIOL PROTEASE THAT RECOGNIZE AND <u>HYDROLYZE</u> A PEPTIDE BOND AT THE C-TERMINAL GLYCINE OF UBIQUITIN.

Biochemical Interactions		
event	P	example
methylate	$\neg\text{methylated}(s)$	<u>METHYLATES</u> PRECORRIN-2 AT THE C-20 POSITION TO PRODUCE PRECORRIN-3A.
phosphorylate	$\neg\text{phosphorylated}(s)$	CAN INTERACT AND <u>PHOSPHORYLATE</u> BUB3.
scavenge	$\neg\text{scavenged}(s)$	PROBABLY SERVES TO <u>SCAVENGE</u> PHOSPHORUS FOR GROWING CELLS.
sever	$\neg\text{severed}(s)$	THESE PROTEINS DO NOT <u>SEVER</u> ACTIN FILAMENTS.
splice	$\neg\text{spliced}(s)$	COULD BE INVOLVED IN <u>SPLICING</u> AND/OR PROCESSING OF CHLOROPLAST RNA'S.

Logical Interactions In line with ([63]), the classification presented here will also distinguish biochemical from logical interactions. In this sense, the *activation* of a certain protein is a logical interaction which can be biochemically realized by different types of interactions such as modifications, cleavages, etc. Certainly, it will not always be straightforward to decide whether an interaction has to be considered as a biochemical or a logical one. In order to classify an event as a biochemical interaction it must be possible to uniquely map it to a corresponding 'low-level' biochemical interaction such as a phosphorylation, a methylation, a cleavage, etc. A logical interaction on the other hand has to be considered as a meta-interaction which, as mentioned above, can be brought about by various 'low-level' interactions. The formalization of such a logical interaction will be almost identical to the one of a biochemical interaction with the only exception that the property changed is in this case a logical property of the protein in question:

$$\begin{aligned}
& \forall e, s_1, s_2 ((e : \text{logical_interaction}(s_1, s_2) \wedge \\
& \text{chemical_substance}(s_1) \wedge \text{chemical_substance}(s_2)) \Leftrightarrow \\
& \exists P, s, s' (\text{logical_property}(P) \wedge (s : \neg P(s_1) \wedge s' : P(s_1) \vee \\
& s : \neg P(s_2) \wedge s' : P(s_2)) \wedge s \supset e \wedge \text{Result}(e, s'))) \quad (28)
\end{aligned}$$

In this sense, the catalytic activation or inhibition of a certain protein can be seen as a special case of such a logical interaction where the corresponding logical property is being catalytically active or not:

$$\begin{aligned}
& \forall e, p_1, p_2 ((e : \text{catalytic_activation}(p_1, p_2) \wedge \text{protein}(p_1) \wedge \text{protein}(p_2)) \Leftrightarrow \\
& \exists s, s' (s : \neg \text{catalytically_active}(p_2) \wedge s \supset e \wedge \\
& s' : \text{catalytically_active}(p_2) \wedge \text{Result}(e, s'))) \quad (29)
\end{aligned}$$

$$\forall e, p_1, p_2 ((e : \text{catalytic_inhibition}(p_1, p_2) \wedge \text{protein}(p_1) \wedge \text{protein}(p_2)) \Leftrightarrow$$

$$\begin{aligned} \exists s, s' (s : \text{catalytically_active}(p_1) \wedge s \supset e \wedge \\ s' : \neg \text{catalytically_active}(p_2) \wedge \text{Result}(e, s')) \end{aligned} \quad (30)$$

Here follow some verbs corresponding to logical interactions as introduced and formalized above. Again the changed logical property is given in the middle column:

Logical Interactions		
event	P	example
activate	$\neg \text{catalytically_active}(s)$	<u>ACTIVATES</u> ARCA BY PHOSPHORYLATION.
destabilize	$\text{stabilized}(s)$	FUNCTION BY <u>DESTABILIZING</u> AN RNA SECONDARY STRUCTURE ON THE 3'END OF THE QBETA PHAGE.
inactivate	$\text{catalytically_active}(s)$	BINDS CRF AND <u>INACTIVATES</u> IT.
inhibit	$\text{catalytically_active}(s)$	KNOWN TO <u>INHIBIT</u> CATHEPSIN B, H, AND L.
open	$\neg \text{open}(s)$	THIS PROTEIN CAN BE ACTIVATED BY CYCLIC GMP WHICH LEADS TO A <u>OPENING</u> OF THE CATION CHANNEL [...]
organize	$\neg \text{structured}(s)$	MAY <u>ORGANIZE</u> MICROTUBULES BY MEDIATING SPINDLE POSITIONING AND MOVEMENT IN THE BUDDING PROCESS.
stabilize	$\neg \text{stabilized}(s)$	THE EXACT FUNCTION OF MAP2 IS UNKNOWN BUT MAPS MAY <u>STABILIZE</u> THE MICROTUBULES AGAINST DEPOLYMERIZATION.
wrap	$\neg \text{wrapped}(s)$	THIS PROTEIN BELONGS TO THE HISTONE LIKE FAMILY OF PROKARYOTIC DNA-BINDING PROTEINS WHICH ARE CAPABLE OF <u>WRAPPING</u> DNA TO STABILIZE IT, [...]

Biochemical Processes Within molecular biology there exist many well-defined special processes which change the property of a biochemical object. The expression of a gene for example changes the property of a gene from being untranscribed to being transcribed. A biochemical process e can thus be formalized as changing a property of its target b :

$$\begin{aligned} \forall e, b ((e : \text{biochemical_process}(b) \wedge \text{biochemical_object}(b)) \Leftrightarrow \\ \exists P, s, s' (\text{process_property}(P) \wedge \\ s : \neg P(b) \wedge s \supset e \wedge s' : P(b) \wedge \text{Result}(e, s'))) \end{aligned} \quad (31)$$

Here are some examples of verbs representing such well-defined processes:

Biochemical Processes		
event	P	example
differentiate	$\neg differentiated(b)$	THIS PROTEIN MIGHT BE INVOLVED IN DETERMINATION AND/OR <u>DIFFERENTIATION</u> OF NERVE CELLS [...]
elongate	$\neg elongated(b)$	INVOLVED IN VIRAL REPLICATION, POSSIBLY IN THE ELONGATION OF DNA.
express	$\neg expressed(b)$	ACTIVATES THE <u>EXPRESSION</u> OF DRUG METABOLIZING ENZYMES GENES (SUCH AS THE CYP1A1 GENE).
pattern	$\neg patterned(b)$	COULD BE INVOLVED IN THE DIFFERENTIATION OF ECTODERMAL LINEAGES AND SUBSEQUENT <u>PATTERNING</u> OF THE EMBRYO.
proliferate	$\neg proliferated(b)$	PROMOTES AG-DEPENDENT <u>PROLIFERATION</u> OF B-CELLS, [...]
replicate	$\neg replicated(b)$	IN THE EARLY STAGE OF INFECTION, T3 DNA <u>REPLICATES</u> AS A LINEAR MONOMER.
transcribe	$\neg transcribed(b)$	RNA POLYMERASE I IS ESSENTIALLY USED TO <u>TRANSCRIBE</u> RIBOSOMAL DNA UNITS.
unwind	$\neg unwound(s)$	CAN <u>UNWIND</u> DOUBLE-STRANDED RNA (HELICASE) AND CAN FOLD OR INTRODUCE A SECONDARY STRUCTURE TO A SINGLE-STRANDED RNA (FOLDASE).

Bind/Dissociate Chemical substances can bind to each other so that as a result they are bound together:

$$\begin{aligned}
 & \forall e, s_1, s_2 ((e : bind(s_1, s_2) \wedge chemical_substance(s_1) \wedge \\
 & \quad chemical_substance(s_2)) \Leftrightarrow \exists s, s' \\
 & \quad (s : \neg bound(s_1, s_2) \wedge s \supset e \wedge s' : bound(s_1, s_2) \wedge Result(e, s'))) \quad (32)
 \end{aligned}$$

The following table contains verbs representing binding events as formalized above:

Binding	
event	example
aggregate	COMPONENT OF THE BASAL LAMINA THAT CAUSES THE <u>AGGREGATION</u> OF ACETYLCHOLINE RECEPTORS AND ACETYLCHOLINE-ESTERASE ON THE SURFACE OF MUSCLE FIBERS OF THE NEUROMUSCULAR JUNCTION.
anchor	MAY <u>ANCHOR</u> THE KINASE TO CYTOSKELETAL AND/OR ORGANELLE- ASSOCIATED PROTEINS, [...]
associate	MAY <u>ASSOCIATE</u> WITH CD21.
bind	MAY <u>BIND</u> A CADHERIN AND PARTICIPATE IN THE TRANSMISSION OF DEVELOPMENTAL INFORMATION.
complex	THIS PROTEIN <u>COMPLEXES</u> WITH CYCLIC AMP AND BINDS TO SPECIFIC DNA SITES NEAR THE PROMOTER TO REGULATE THE TRANSCRIPTION OF SEVERAL <u>CATABOLITE-SENSITIVE OPERONS</u> .
dock	PROMOTES <u>DOCKING</u> OF IMPORT SUBSTRATES TO THE <u>NUCLEAR ENVELOPE</u> .
link	MAY <u>LINK</u> THE MECHANOSENSORY CHANNEL AND THE MICROTUBULE CYTOSKELETON OF THE TOUCH RE-CEPTOR NEURONS.

On the other hand, there are also events which separate substances from each other so that they are not bound anymore:

$$\begin{aligned} &\forall e, s_1, s_2 ((e : dissociate(s_1, s_2) \wedge chemical_substance(s_1) \wedge \\ &chemical_substance(s_2)) \Leftrightarrow \exists s, s' \\ &(s : bound(s_1, s_2) \wedge s \supset e \wedge s' : \neg bound(s_1, s_2) \wedge Result(e, s'))) \quad (33) \end{aligned}$$

Dissociation	
event	example
dissociate	A49 IS EASILY <u>DISSOCIATED</u> FROM THE REST OF POL A, PRODUCING THE FORM A*, [...]
release	PLAYS A ROLE IN THE <u>RELEASE</u> OF TFIIB FROM THE TRANSCRIPTION COMPLEX DURING TRANSCRIPTION INITIATION.
separate	AT THE NUCLEOPLASMIC SIDE OF THE NPC, THE THREE COMPONENTS <u>SEPARATE</u> AND IMPORTIN-ALPHA AND -BETA ARE RE- EXPORTED FROM THE NUCLEUS TO THE CYTOPLASM.

Formation Some biochemical processes form new substances, complexes or structures. In order to formalize such a process, a notion of physical existence - as opposed to logical existence as denoted by the existence quantifier \exists - has to be introduced. The predicate $E(x)$ will be used for this purpose. In this line, an event in which something is formed can be formalized as follows ¹³:

¹³The meaning of a formation as considered here corresponds to the first one given by Reyle et al. in [59].

$$\begin{aligned} \forall e, b ((e : formation(b) \wedge biochemical_object(b)) \Leftrightarrow \\ \exists s, s' (s : \neg E(b) \wedge s \supset e \wedge s' : E(b) \wedge Result(e, s'))) \end{aligned} \quad (34)$$

Again, here follow some examples:

Formation	
event	example
create	ACTS BY DELIVERING COPPER TO <u>CREATE</u> FUNCTIONAL HORMONE RECEPTORS
develop	MAY ALSO PARTICIPATE IN THE <u>DEVELOPMENT</u> OF THE CELLS OF THE INNER NUCLEAR LAYER, [...]
establish	REQUIRED FOR THE <u>ESTABLISHMENT</u> OF STABLE CONNECTIONS BETWEEN THE LARVAL OPTIC NERVES, [...]
form	REQUIRED FOR MEIOTIC RECOMBINATION, SYNAPTONEMAL COMPLEX <u>FORMATION</u> AND CELL CYCLE PROGRESSION.
give	CATALYZES THE CONDENSATION OF TWO MOLECULES OF GLYOXYLATE TO <u>GIVE</u> 2-HYDROXY-3-OXOPROANOATE.
originate	APPEARS TO PERFORM A VERY EARLY FUNCTION IN ESTABLISHING THE IDENTITY OF A SUBSET OF CELLS THAT <u>ORIGINATE</u> IN THE REGION OF THE VENTRICULAR ZONE IN THE DEVELOPING SPINAL CHORD AND IN THE HINDBRAIN.
produce	THIS ENZYME IS REQUIRED BY THE INSECT IMMUNE SYSTEM TO <u>PRODUCE</u> MELANIN WHICH IS USED TO ENGULF FOREIGN OBJECTS.
synthesize	IT CAN <u>SYNTHESIZE</u> BOTH PRIMARY AND SECONDARY ALCOHOL ESTERS.
generate	COMPONENT OF RIBONUCLEASE P, A PROTEIN COMPLEX THAT <u>GENERATES</u> MATURE TRNA MOLECULES BY CLEAVING THEIR 5' ENDS.

Integrity There are many biochemical processes which affect the integrity of a protein, a chemical substance or a source such as an organism, a cell, a membrane, etc. The fact that the integrity of something has not been affected will be captured by the predicate *intact(b)*. Thus an event which affects the physical integrity of a biochemical object in a negative manner can be formalized as follows:

$$\begin{aligned} \forall e, b ((e : disturb_integrity(b) \wedge biochemical_object(b)) \Leftrightarrow \\ \exists s, s' (s : intact(b) \wedge s \supset e \wedge s' : \neg intact(b) \wedge Result(e, s'))) \end{aligned} \quad (35)$$

On the other hand, in case the event restores the integrity, the formalization would be as follows:

$$\begin{aligned} \forall e, b ((e : \text{restore_integrity}(b) \wedge \text{biochemical_object}(b)) \Leftrightarrow \\ \exists s, s' (s : \neg \text{intact}(b) \wedge s \supset e \wedge s' : \text{intact}(b) \wedge \text{Res}(e, s'))) \end{aligned} \quad (36)$$

The following table contains verbs representing events which affect the integrity of a biochemical object. The column in the middle indicates whether it is an event which disturbs or restores the integrity:

Integrity		
event	type	example
destroy	disturb	G2/M CYCLINS ACCUMULATE STEADILY DURING G2 AND ARE ABRUPTLY DESTROYED AT MITOSIS.
heal	restore	PLAYS AN IMPORTANT ROLE IN STIMULATING ADJACENT CELLS TO GROW AND THEREBY HEAL THE WOUND.
infect	disturb	PROTECTS AGAINST <u>INFECTION</u> BY A NEUROVIRULENT STRAIN OF SINDBIS VIRUS.
regenerate	restore	THE COMPONENT A MAY BE REGENERATED BY TRANSFERRING ITS PRENYLATED RAB TO A PROTEIN ACCEPTOR.

Availability Many biochemical processes act as an intermediate step in a chain of processes making some substance available for some other processes using it as a cofactor or as a catalyst. The fact that a specific substance is available will be expressed through the predicate *available(c)*. Thus, events which make some substance available can be formalized as follows:

$$\begin{aligned} \forall e, c ((e : \text{make_available}(c) \wedge \text{chemical_substance}(c)) \Leftrightarrow \exists s, s' \\ (s : \neg \text{available}(c) \wedge s \supset e \wedge s' : \text{available}(c) \wedge \text{Result}(e, s'))) \end{aligned} \quad (37)$$

Here follows a list of different events which make some substance available together with an example:

Availability	
event	example
present	ADHESION MOLECULE THAT ACCOMPLISHES CELL BINDING BY <u>PRESENTING</u> CARBOHYDRATE(S) TO THE LECTIN DOMAIN OF L-SELECTIN.
procure	MAY PLAY A ROLE IN ALL ASPECTS OF HEMOGLOBIN PRODUCTION: GLOBIN SYNTHESIS, HEME SYNTHESIS, AND THE <u>PROCUREMENT</u> OF IRON.
provide	PROPOSED TO <u>PROVIDE</u> ACTIVATED SULFATE FOR TRANSFER TO NOD FACTOR.
recycle	MAY ALSO CLEAVE INTRACELLULARLY GENERATED PEPTIDES TO <u>RECYCLE</u> AMINO ACIDS FOR PROTEIN SYNTHESIS.

Change of Location Some biochemical processes are responsible for transporting specific substances from one location to another. In the model presented here, every chemical substance c will have a certain location l assigned by the predicate $location(c, l)$. Thus, every change of location can be formalized as follows:

$$\begin{aligned} \forall e, c ((e : change_location(c) \wedge chemical_substance(c)) \Leftrightarrow \\ \exists s, s' (s : location(c, l_1) \wedge s \supset c \wedge \\ s' : location(c, l_2) \wedge l_1 \neq l_2 \wedge Result(e, s'))) \end{aligned} \quad (38)$$

The following table contains verbs representing change of location events. The special pre- and postconditions introduced by each event can be found in the middle column:

Change of Location		
event	special condition	example
attract		MAY BE INVOLVED IN FORMATION AND FUNCTION OF THE MUCOSAL LYMPHOID TISSUES BY <u>ATTRACTING</u> LYMPHOCYTES AND DENDRITIC CELLS TOWARDS EPITHELIAL CELLS.
bring		MAY HAVE A DIRECT ROLE IN ACTIVATION OF RAC AND/OR RHO AND IN <u>BRINGING</u> THE ACTIVATED GTPASE TO SPECIFIC TARGET SITES SUCH AS MICROTUBULES.
deliver		SEEM TO FUNCTION AS A FUSION PROTEIN REQUIRED FOR THE <u>DELIVERY</u> OF CARGO PROTEINS TO ALL COMPARTMENTS OF THE GOLGI STACK INDEPENDENT OF VESICLE ORIGIN.
direct		MAY <u>DIRECT</u> THIOLEASE TO PEROXISOMES BY SHUTTLING BETWEEN THE CYTOSOL AND PEROXISOMAL MEMBRANES.
export	$l_2 \neq nucleus$	MAY BE INVOLVED IN <u>EXPORT</u> OF PERIPLASMIC FLAGELLA PROTEINS.
internalize	$l_2 = nucleus$	IN ORDER TO BE <u>INTERNALIZED</u> , THE RECEPTOR-LIGAND COMPLEXES MUST FIRST CLUSTER INTO CLATHRIN-COATED PITS.
sequesterate		PROBABLY INVOLVED IN ZINC TRANSPORT OUT OF THE CYTOPLASM, MAY BE BY <u>SEQUESTRATION</u> INTO AN INTRACELLULAR COMPARTMENT.

Change of Location		
event	special condition	example
repartition		MAY PLAY A ROLE IN <u>REPARTITIONING</u> CALCIUM BETWEEN THE TWO MAJOR INTRACELLULAR CALCIUM STORES IN ASSOCIATION WITH THE 19 KDA OR BCL-2 PROTEINS.
shuttle		MAY DIRECT THIOLASE TO PEROXISOMES BY <u>SHUTTLLING</u> BETWEEN THE CYTOSOL AND PEROXISOMAL MEMBRANES.
recruit		MIGHT CHANGE THE DNA-BINDING SPECIFICITY OF OTHER TRANSCRIPTION FACTORS AND <u>RECRUIT</u> THEM TO UNUSUAL DNA SITES.
target		BINDS TO TYPE II REGULATORY SUBUNITS OF PROTEIN KINASE A AND <u>ANCHORS/TARGETS</u> THEM TO THE NUCLEAR MEMBRANE OR SARCOPLASMIC RETICULUM.
translocate		PLAYS A DIRECT ROLE IN THE <u>TRANSLOCATION</u> OF SOME PRESECRETORY PROTEINS INTO THE ENDOPLASMIC RETICULUM.
transport		<u>TRANSPORTS</u> CALCITONIN-RECEPTOR-LIKE RECEPTOR TO THE PLASMA MEMBRANE WHERE IT ACTS AS A CALCITONIN-GENE-RELATED PEPTIDE RECEPTOR.

Modification of Structure Some events modify the (mereological) structure of some molecules (and thus also their chemical properties) by inserting, removing or replacing some parts of them. In the following it will be formalized what an insertion, a removal and a replacement are:

$$\begin{aligned}
& \forall e, b, c ((e : \textit{insertion}(b, c) \wedge \textit{chemical_substance}(c) \wedge \\
& \textit{biochemical_object}(b)) \Leftrightarrow \exists s, s' \\
& (s : \neg \textit{part_of}(b, c) \wedge s \supset c \wedge s' : \textit{part_of}(b, c) \wedge \textit{Result}(e, s'))) \quad (39)
\end{aligned}$$

$$\begin{aligned}
& \forall e, b, c ((e : \textit{removal}(b, c) \wedge \textit{chemical_substance}(c) \wedge \\
& \textit{biochemical_object}(b)) \Leftrightarrow \exists s, s' \\
& (s : \textit{part_of}(b, c) \wedge s \supset c \wedge s' : \neg \textit{part_of}(b, c) \wedge \textit{Result}(e, s'))) \quad (40)
\end{aligned}$$

$$\begin{aligned}
& \forall e, b_1, b_2, c ((e : \textit{replacement}(b_1, b_2, c) \wedge \textit{chemical_substance}(c) \wedge \\
& \textit{biochemical_object}(b_1) \wedge \textit{biochemical_object}(b_2)) \Leftrightarrow
\end{aligned}$$

$$\begin{aligned} \exists s, s' (s : part_of(b_1, c) \wedge s : \neg part_of(b_2, c) \wedge s \supset e \wedge \\ s' : \neg part_of(b_1, c) \wedge s' : part_of(b_2, c) \wedge Result(e, s')) \end{aligned} \quad (41)$$

A phosphorylation for example can be seen as the insertion of a phosphate group into a protein¹⁴:

$$\begin{aligned} \forall e, r, c ((e : phosphorylation(r, c) \wedge chemical_substance(c) \wedge \\ phosphate_group(r)) \Leftrightarrow \exists s, s' \\ (s : \neg part_of(r, c) \wedge s \supset e \wedge s' : part_of(r, c) \wedge Result(e, s'))) \end{aligned} \quad (42)$$

The verbs in the following table represent insertion/removal/replacement events as formalized by the above formulas. The column in the middle indicates which type of event it corresponds to:

Modification		
event	type	example
exchange	replacement	CATALYZES THE HYDROLYSIS OF ATP COUPLED WITH THE <u>EXCHANGE</u> OF H(+) AND K(+) IONS ACROSS THE PLASMA MEMBRANE.
excise	removal	HYDROLYSIS OF THE DEOXYRIBOSE N-GLYCOSIDIC BOND TO <u>EXCISE</u> 3-METHYLADENINE FROM THE DAMAGED DNA POLYMER FORMED BY ALKYLATION LESIONS.
insert	insertion	MAY <u>INSERT</u> THE NASCENT SPB INTO THE NUCLEAR ENVELOPE.
remove	removal	<u>REMOVES</u> THE AMINO-TERMINAL METHIONINE FROM NASCENT PROTEINS.
replace	replacement	PROMOTES THE ACTIVATION OF ARF THROUGH <u>REPLACEMENT</u> OF GDP WITH GTP.
substitute	replacement	THE HYDROXYL GROUP IS <u>SUBSTITUTED</u> BY A HYDRIDE DERIVED FROM THE NOW REDUCED FAD IN AN SN2' REACTION LEADING TO VINYLACETYL-COA.
transfer	removal	CATALYSES THE MAGNESIUM ION DEPENDENT <u>TRANSFER</u> OF A RIBOSYL PHOSPHATE GROUP FROM PRPP TO THE N9 NITROGEN OF HYPOTHANTHINE, GUANINE OR XANTHINE.

Order Some verbs express the logical or temporal order in which biochemical processes occur. Though these verbs certainly also have an event reading, only the interpretation as constraint on the order of the events in question will be considered here. The table below contains such verbs imposing an order on two

¹⁴This view of a phosphorylation corresponds more or less to the one in [58].

events e and e' . For each verb the specific relation between e and e' is given in the middle column:

Order		
event	temporal_order(e, e')	example
follow	$e < e'$	INDUCES A TRANSIENT PERIOD OF FAST FLICKERING IN THE CHANNEL OPENINGS, <u>FOLLOWED</u> BY AN ALMOST COMPLETE BLOCKADE OF THE CHANNEL.
precede	$e > e'$	JUNCTION OF THE 60S RIBOSOMAL SUBUNIT TO FORM THE 80S INITIATION COMPLEX IS <u>PRECEDED</u> BY HYDROLYSIS OF THE GTP BOUND TO EIF-2 AND RELEASE OF AN EIF-2-GDP BINARY COMPLEX.

5.2.3 The Realistic Picture

The rather impressionistic picture of protein interactions and their logical relations developed in section 5.1 has been turned into a more realistic and detailed picture of these interactions by assigning them clear semantics. This has been achieved by indicating the pre- and postconditions which describe the change brought about by a specific event or interaction. These pre- and postconditions have been formalized with a minimal number of predicates capturing the value v of a certain property P of a biochemical process bp ($value(P(bp), v)$), the fact that a biochemical object exists ($E(b)$), that two substances s_1 and s_2 are bound together ($bound(s_1, s_2)$), that a biochemical object b is intact ($intact(b)$) or that a chemical substance s is available ($available(s)$), the fact that every substance s has a certain location l within the cell ($location(s, l)$), that a biochemical object b is part of a certain mereological structure s ($part_of(b, s)$) as well as the relations expressing the chronological order between events.

5.3 Design Principles

After having developed a model of the general interactions between proteins in section 5.1 and given a formal semantics to some of these interactions in section 5.2, it has to be clarified how these two 'pictures' can be combined to yield an ontology as defined in section 4.2. In particular it has to be clarified how events will be represented as concepts of such an ontology. Furthermore, the principles which allow to link two concepts together by a *is-a* relation or some other relation such as *Result*, *Explanation* or *Elaboration* have to be discussed.

5.3.1 Ontological Representation of Events

As mentioned before, the basic concepts (or classes) of the ontology are events. So, the question of how to represent them consequently becomes a central issue

within the design of such an ontology.

As the purpose of the ontology is the resolution of anaphora, especially concentrating on events as antecedents, it seems intuitive to use a representation of events in a Davidsonian tradition ([14]) by introducing variables for them such that they can be referred to later on in discourse.

However, this doesn't answer the question how to treat the roles that constitute an event. Roles will be understood as entities which describe a certain aspect of an event and thus are related to it. In particular, the agent, the patient, the beneficiary or the instrument of an event are such roles.

In this section, two different approaches will be presented and discussed. The first one will consist in the explicit representation of all the roles involved within an event as arguments of the predicate describing it. This approach will be called the 'frame-based approach'. In order to illustrate the different approaches the binding of two proteins as a representative example event will be discussed. The following simplified natural language definition of the binding of two proteins will be used as starting point of this formalization discussion:

Definition 19

A protein-protein binding event e consists in a protein p_1 binding a protein p_2 at location d_2 through location d_1 with a certain affinity a .

The frame-based representation will now look as follows:

$$\begin{aligned} \exists e, p_1, p_2, d_1, d_2, a \ (bind(e, p_1, d_1, p_2, d_2, a) \wedge \\ event(e) \wedge protein(p_1) \wedge protein(p_2) \wedge \\ domain(d_1) \wedge domain(d_2) \wedge affinity(a)) \end{aligned} \quad (43)$$

The alternative representation of an event follows the Davidsonian tradition of representing prepositional modifiers (as well as other modifiers) as related in some way to the event. This way of representing events will be referred to as the 'Davidsonian approach'. In this Davidsonian approach, the protein binding event would have the following representation:

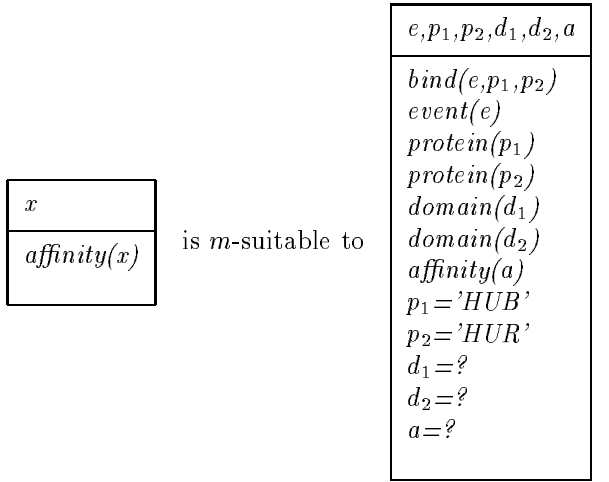
$$\begin{aligned} \exists e, p_1, p_2, d_1, d_2, a \ (bind(e, p_1, p_2) \wedge \\ event(e) \wedge protein(p_1) \wedge protein(p_2) \wedge \\ domain(d_1) \wedge domain(d_2) \wedge affinity(a) \wedge \\ binding_domain(e, p_1, d_1) \wedge binding_domain(e, p_2, d_2) \wedge \\ binding_affinity(e, a)) \end{aligned} \quad (44)$$

Before going ahead with the discussion of both approaches, it should be first mentioned that it is very rare to find all the roles conceptually constituting a certain event explicitly stated within a sentence or utterance. This observation has different consequences for both of the approaches mentioned. The consequence for the frame-base representation is that most of the arguments of the

corresponding event predicate will be underspecified while in the Davidsonian approach they will be simply omitted. On the other hand, for anaphora resolution it is necessary that omitted arguments are made explicit in order to resolve an anaphoric expression as in the following (made up) example:

(45) HUB binds HUR. The affinity is high.

This resolution poses no problem for the frame-based representation as we can, informally speaking, identify the affinity mentioned in the second sentence with the underspecified affinity of the binding event in the first sentence thus resolving the anaphora by linking. Formalized in DRT, it is obvious that



In order to make the omitted information explicit within the Davidsonian approach, it will need to be inferred:

$$\begin{aligned}
& \forall e, p_1, p_2 ((bind(e, p_1, p_2) \wedge event(e) \wedge protein(p_1) \wedge protein(p_2)) \rightarrow \\
& \exists d_1, d_2, a : (domain(d_1) \wedge domain(d_2) \wedge affinity(a) \wedge \\
& binding_domain(e, p_1, d_1) \wedge binding_domain(e, p_2, d_2) \wedge \\
& binding_affinity(e, a))) \tag{46}
\end{aligned}$$

In fact, the anaphora will also get resolved but by bridging instead of linking as the anaphoric expression refers to an inferred entity rather than to an entity explicitly stated as omitted as in the frame-based approach. Thus, the choice of one of the two approaches has not only consequences for the way omitted roles are made explicit but also on the the way anaphoric expressions are resolved. Let's now discuss the following (also made up) example:

(47) HUB binds HUR at HELIX 1. HUB also binds HUR at STEM LOOP I.

Obviously, in example (47) both events are not coreferring as the stated domains are incompatible. The frame-based representation allows such an analysis in a straightforward way as the incompatibility between the domains leads to the

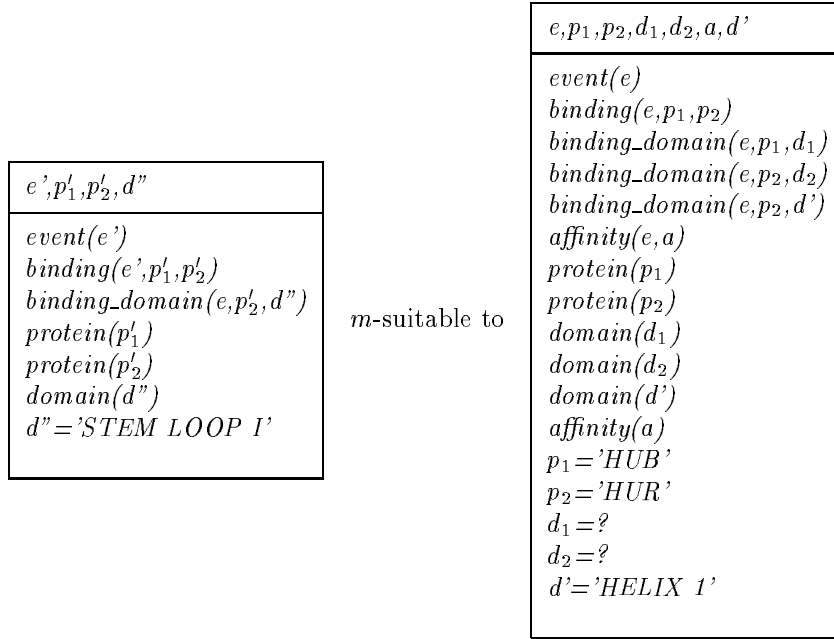
fact that

e, p_1, p_2, d_1, d_2, a $bind(e, p_1, p_2)$ $event(e)$ $protein(p_1)$ $protein(p_2)$ $domain(d_1)$ $domain(d_2)$ $affinity(a)$ $p_1 = 'HUB'$ $p_2 = 'HUR'$ $d_1 = ?$ $d_2 = 'HELIX 1'$ $a = ?$	is not m -suitable to	e, p_1, p_2, d_1, d_2, a $bind(e, p_1, p_2)$ $event(e)$ $protein(p_1)$ $protein(p_2)$ $domain(d_1)$ $domain(d_2)$ $affinity(a)$ $p_1 = 'HUB'$ $p_2 = 'HUR'$ $d_1 = ?$ $d_2 = 'STEM LOOP I'$ $a = ?$	
---	-------------------------	---	--

However, things turn out to be a little bit more complicated when sticking to the Davidsonian representation. In fact, if we represent the first sentence in example (47) as

$$\begin{aligned}
 & \exists e, p_1, p_2, d_2 (bind(e, p_1, p_2) \wedge binding_domain(e, p_2, d')) \\
 & protein(p_1) \wedge protein(p_2) \wedge domain(d') \\
 & p_1 = 'HUB' \wedge p_2 = 'HUR' \wedge d' = 'HELIX 1') \tag{48}
 \end{aligned}$$

because of axiom (46) we will get that



So the problem is that there seems to be no straightforward way to associate the binding domain d' explicitly stated in the sentence with the implicitly inferred domain d_2 which is underspecified and thus represents a suitable antecedent for any domain. In order to solve this problem there are basically two possibilities. As a first solution, we could avoid the coreferring reading by avoiding that the existence of a domain is inferred. Thus we would need to overwrite axiom (46) (by a certain type of penguin principle) with the more specific axiom:

$$\begin{aligned} & \forall e, p_1, p_2, d_2 ((bind(e, p_1, p_2) \wedge binding_domain(e, p_2, d_2)) \rightarrow \\ & \exists d_1, a(binding_domain(e, p_1, d_1) \wedge affinity(e, a))) \end{aligned} \quad (49)$$

For the protein binding case discussed this would thus yield $2^3 = 8$ specific rules overwriting the general axiom (46) (one for each combination of additional predicates characterizing the binding). As an alternative solution we could introduce a certain type of uniqueness condition on the domain explicitly stated:

$$\begin{aligned} & \forall e, p_1, p_2, d, d' ((bind(e, p_1, p_2) \wedge binding_domain(e, p_2, d) \wedge \\ & binding_domain(e, p_2, d')) \rightarrow d' = d) \end{aligned} \quad (50)$$

Thus, the coreferring reading will be ruled out by logical inconsistency. In general, both solutions seem unsatisfactory as they either lead to a combinatorial explosion of specific rules or to the necessity of consistency checking within the anaphora resolution process.

The discussion has shown that a frame-based representation seems more suitable

for the anaphora resolution process. The consequence of the use of such a representation is that examples as the following one will be resolved via linking and not via bridging, which is reasonable because both sentences refer to one and the same event entity:

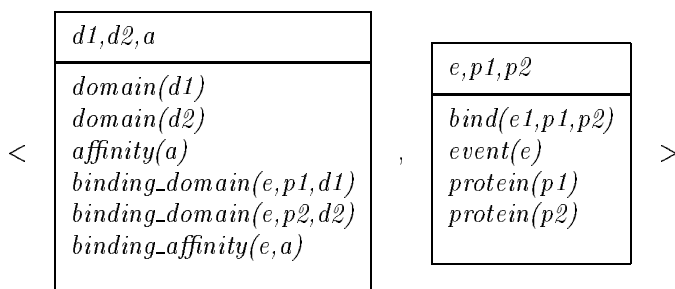
(51) HUB binds HUR. It binds HUR with high affinity.

A further consequence of a frame-based representation is that the value of most of the discourse referents will be underspecified, but it is not clear if this has to be seen as a disadvantage. In fact, the example above shows that underspecified referents can and will be filled further on in the discourse. Furthermore, anaphora resolution requires that omitted information is made explicit such that discourse referents marked as underspecified seem to be unavoidable at least at the DRS level. However, one may think of eliminating such underspecified discourse referents when embedding the DRS into a certain logical model, but it should be noted that this idea is very speculative and will not be further developed as it would reach far beyond the aim of anaphora resolution pursued within this work.

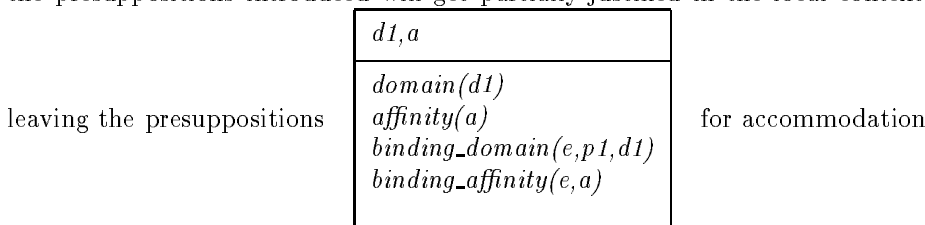
A certain disadvantage is that by putting all the logical arguments of an event into a predicate representing it, the logical relations between them will be lost. But there's an intuitive solution to this. We just have to extend the frame-based representation making the relation between an argument and the event explicit. In this sense, the representation of a binding between two proteins can be extended as follows:

$$\begin{aligned}
& \exists e, p_1, p_2, d_1, d_2, a (bind(e, p_1, d_1, p_2, d_2, a) \wedge \\
& event(e) \wedge protein(p_1) \wedge protein(p_2) \wedge \\
& domain(d_1) \wedge domain(d_2) \wedge affinity(a) \wedge \\
& binding_domain(e, p_1, d_1) \wedge binding_domain(e, p_2, d_2) \wedge \\
& binding_affinity(e, a))
\end{aligned} \tag{52}$$

Now that both approaches have been discussed to some extent, a third one should be mentioned and discussed. This third approach consists in making implicit arguments explicit by introducing presuppositions for each event ([28], [60], [25]). According to presupposition theory, presuppositions introduced by such means have to be either satisfied by the context or accommodated or - as Kamp et al. ([28]) prefer - 'justified' meaning that presupposition satisfaction is a combination of (partial) satisfaction and (partial) accommodation. Following this line, we could model events in a DRT style manner (following [25]) as introducing presuppositions. The entry for the binding event (19) would thus look as follows:



So presupposition computation (as Kamp ([25]) calls it) would in our case be ontology driven in the sense that the axiom defining what a binding is should describe which presuppositions it introduces. But what about presupposition satisfaction/justification? Justification is generally seen as a process which can either take place at the local (i.e. the processed sentence or phrase) or at the global context. In this line, if we encounter a sentence like the first one in example (47), it would get a Davidsonian representation as in (48) such that the presuppositions introduced will get partially justified in the local context



thus making omitted arguments again explicit such that they can be referred to later in discourse. This is exactly what we want, so why not modeling implicit arguments via presuppositions?

First of all we are not dealing with complex linguistic phenomena like the disambiguation of the senses of 'wieder' (Engl. 'again') as in ([28]) or the disambiguation of temporal underspecification in discourse via presupposition justification ([60]), such that the use of presuppositions might embed the approach within the sound framework of presupposition theory but possibly with too much effort and too less effective benefit resulting from this. Furthermore, the use of presuppositions would lead to the study and implementation of presupposition satisfaction and this would carry far way out of the scope of the work presented here. As a last and probably most important argument it should be pointed out that an application-independent ontology will be used as source of world knowledge driving the anaphora resolution process and it can not be expected of such an ontology to have implicit arguments of events modeled in the linguistically motivated form of presuppositions.

5.3.2 Taxonomic Relations

In this section it will be explained how the semantics of events as defined in section 5.2 as well as a hierarchical taxonomy of their arguments can play an

important role in ontology design and in particular in the development of a hierarchical taxonomy of events. In section 5.3.1, for example, a stimulation-event has been defined as a special case of a control/regulation-event in the sense that the value of some property of the biochemical process in question is increased, i.e. $v_2 > v_1$. Thus, it is the semantics of both events which allows us to represent a stimulation-event as a special case of a control/regulation and to add the following taxonomic relation to the set A of the ontology:

$$\begin{aligned} \forall e, p, bp ((e : \textit{stimulation}(p, bp) \wedge \textit{protein}(p) \wedge \textit{biochemical_process}(bp)) \rightarrow \\ e : \textit{control/regulation}(p_1, bp)) \end{aligned} \quad (53)$$

Furthermore, the activation of a biochemical process has been defined as a special case of a control/regulation event for which the special condition $v_1 = 0$ holds. Thus, in fact an activation can even be regarded as a special case of a stimulation event:

$$\begin{aligned} \forall e, p, bp ((e : \textit{catalytic_activation}(p, bp) \wedge \textit{protein}(p) \wedge \\ \textit{biochemical_process}(bp)) \rightarrow e : \textit{stimulation}(p, bp)) \end{aligned} \quad (54)$$

On the other hand, the inhibition of a biochemical process is a control/regulation event with the special condition $v_2 = 0$:

$$\begin{aligned} \forall e, p, bp ((e : \textit{inhibition}(p, bp) \wedge \textit{protein}(p) \wedge \textit{biochemical_process}(bp)) \rightarrow \\ e : \textit{control/regulation}(p, bp)) \end{aligned} \quad (55)$$

In section 5.2 the catalytic activation of a protein p_2 by a protein p_1 has been presented as a special case of a logical interaction between proteins assuming that the property of being *catalytic_active* is a logical one, i.e. $\textit{logical_property}(\lambda x. \textit{catalytic_active}(x))$:

$$\begin{aligned} \forall e, p_1, p_2 ((e : \textit{activation}(p_1, p_2) \wedge \textit{protein}(p_1) \wedge \textit{protein}(p_2)) \rightarrow \\ e : \textit{logical_interaction}(p_1, p_2)) \end{aligned} \quad (56)$$

But can the binding of two proteins also be seen as some sort of interaction between them? Indeed this is the case if we regard $\lambda p_1. \textit{bound}(p_1, p_2)$ or $\lambda p_2. \textit{bound}(p_1, p_2)$ as biochemical properties, i.e. $\forall p_1 (\textit{protein}(p_1) \rightarrow \textit{biochemical_property}(\lambda x. \textit{bound}(x, p_1)))$, and thus as specializations of the relation P in definition (27). Thus, the following taxonomic relation can be added to the ontology:

$$\begin{aligned} \forall e, p_1, p_2 ((e : \textit{binding}(p_1, p_2) \wedge \textit{protein}(p_1) \wedge \textit{protein}(p_2)) \rightarrow \\ e : \textit{biochemical_interaction}(p_1, p_2)) \end{aligned} \quad (57)$$

Of course, the same argumentation as above then holds for a *dissociation* event so that the whole *bind/dissociate* class can in fact be seen as a specialization of the *biochemical_interaction* class. So it is obvious that the classes proposed in section 5.2.2 are not disjoint. The reason is that the classes are linguistically

motivated in the sense that the aim of the classification was to group as much verbs as possible into a reasonable small number of classes and then give an appropriate and general semantics for each of these classes. If the aim had been to yield a disjoint classification of events, the proceeding would have had to be the other way round, i.e. first a number of conceptual and disjoint classes would have had to be identified and then suitable verbs would have had to be found for each class.

Further we could ask if all the classes are related to each other, but obviously the answer to this question depends on how the classes *biochemical_property*, *logical_property* and *process_property* are defined. If they are disjoint then the classes *biochemical_interaction*, *logical_interaction* and *biochemical_process* will in consequence also be disjoint. Within this work these classes will in fact be regarded as disjoint and the class *biochemical_event* (compare figure 1) will be defined as the disjoint union of the classes *biochemical_interaction*, *logical_interaction*, *control/regulation* and *biochemical_process*. Furthermore, the physical existence ($E(b)$) as well as the location ($location(b, l)$) of a biochemical object will be seen as logical properties of it. The property of being part of a larger mereological structure ($part_of(b, s)$) will be seen as a biochemical one. Consequently, the formation of a biochemical object as well as its transport will both be logical interactions while a modification will be a biochemical one (compare figure 1). Finally, it should be explained how a taxonomic hierarchy of the arguments of events can in turn contribute to a hierarchical classification of events. For example the fact that a DNA as well as a RNA molecule are special kinds of nucleic acids allows to identify a binding of a protein to a DNA molecule or a RNA molecule as special cases of the binding of a protein to a nucleic acid:

$$\begin{aligned} \forall e, p_1, p_2 ((e : binding(p_1, dna) \wedge protein(p_1) \wedge DNA(dna)) \rightarrow \\ e : binding(p_1, dna) \wedge nucleic_acid(dna)) \end{aligned} \quad (58)$$

$$\begin{aligned} \forall e, p_1, rna ((e : binding(p_1, rna) \wedge protein(p_1) \wedge RNA(rna)) \rightarrow \\ e : binding(p_1, rna) \wedge nucleic_acid(rna)) \end{aligned} \quad (59)$$

Figure (3) shows that the possibility of generalizing over the arguments of a certain event also introduces the possibility of multiple inheritance into the ontology such that the classes of the ontology are in general not disjoint. However, the author is not aware of any disadvantage this may lead to.

The obvious advantage of defining such a taxonomic hierarchy of events, which from a mathematical point of view is a partial order as defined in section 4.2, is that conceptual relations between different events can be defined as high in the hierarchy as possible so that concepts which are lower in the hierarchy and thus more special also “inherit” this conceptual relation. This makes the ontology more transparent as well as easier to develop, understand and maintain by reducing the number of conceptual definitions in the set D.

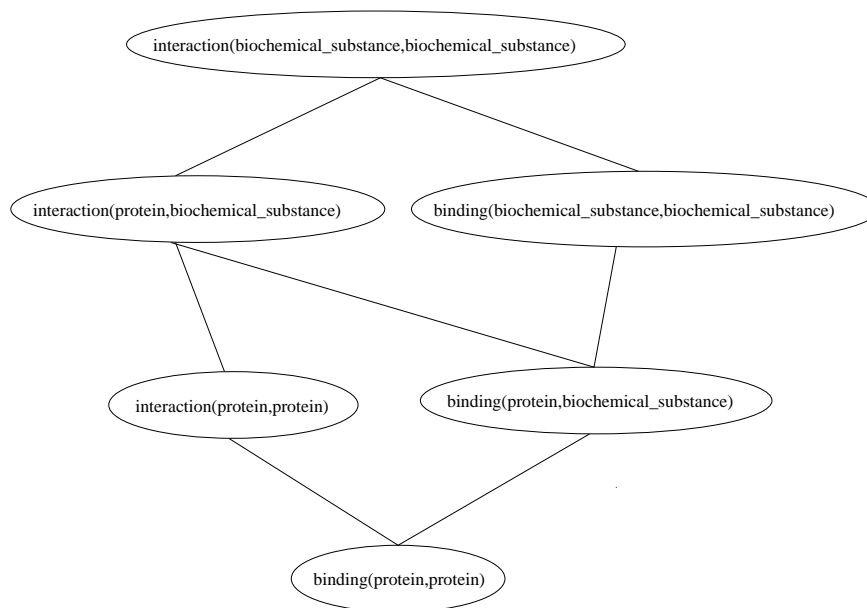


Figure 3: Multiple inheritance in the ontology

5.3.3 Conceptual Definitions

Before getting into details concerning the principles and methods used to model the relevant conceptual definitions of the domain in question, it should be first explained what a conceptual definition is supposed to be. The view of conceptual definitions underlying the work presented here is that they describe the nature of a certain concept or object by specifying other concepts or objects which are related to it in a specific way. As mentioned in section 4.2 it is in the nature of rooms to have a lamp, walls and a ceiling as parts. When applying this idea to the domain of molecular biology we could say that it is in the nature of a binding between two proteins to produce a complex as a result. Similarly, from an 'overall perspective', it could be claimed that it is in the nature of protein interaction to regulate some other biochemical processes as a result.

After these introductory comments, it will now be shown how some of the edges of figure 1 in section 5.1 can be translated into logical axioms representing conceptual definitions as defined above. Let's for example take the edge stating that a protein interaction results in a control/regulation of a biochemical process. This fact can be represented by the following axiom:

$$\forall e, p_1, p_2 ((e : \text{biochemical_interaction}(p_1, p_2) \wedge \text{protein}(p_1) \wedge \text{protein}(p_2)) \rightarrow$$

$$\begin{aligned} \exists bp, e' (e' : control/regulation(p_1, bp) \wedge biochemical_process(bp) \wedge \\ Result(e, e')) \end{aligned} \quad (60)$$

On the other hand the fact that a control/regulation event presupposes a certain interaction will be formalized by inferring the existence of a biochemical interaction which explains it:

$$\begin{aligned} \forall e, p_1, bp ((e : control/regulation(p_1, bp) \wedge protein(p_1) \wedge \\ biochemical_process(bp)) \rightarrow \exists e', c \\ (e' : biochemical_interaction(p_1, c) \wedge chemical_substance(c) \wedge \\ Explanation(e, e'))) \end{aligned} \quad (61)$$

Similarly, the fact that the modification of a protein presupposes a binding can be formalized as follows:

$$\begin{aligned} \forall e, p_1, p_2 ((e : modification(p_1, p_2) \wedge protein(p_1) \wedge protein(p_2)) \rightarrow \\ \exists e' (e' : binding(p_1, p_2) \wedge Explanation(e, e'))) \end{aligned} \quad (62)$$

The examples show that given a suitable representation of the events of the domain in question as well as a set of logical relations such as the ones defined by Lascarides et al. ([35]), the conceptual relations between different events can be expressed in form of logical axioms. These logical axioms will then constitute the set D of conceptual definitions of the ontology.

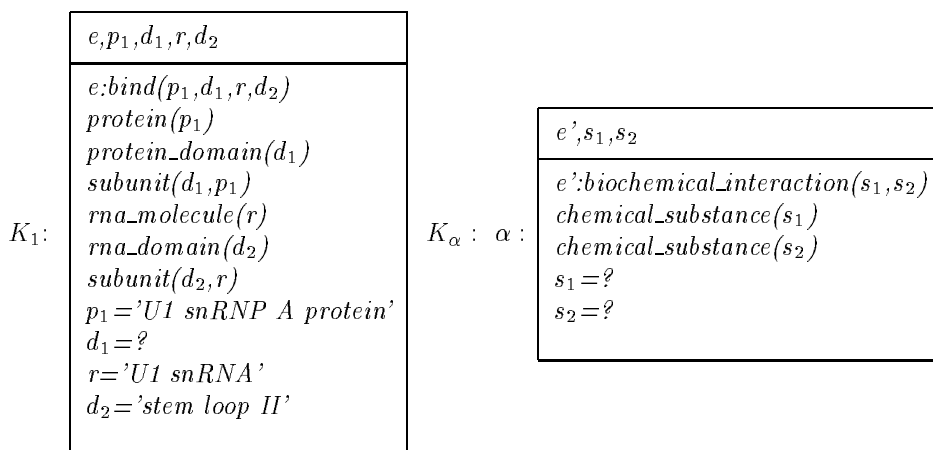
6 Application to Examples

In this section the ontology-driven approach to bridging reference resolution will be applied to some examples of the SWISS-PROT corpus. In the discussion of these examples, it will be assumed that a semantic representation for the relevant events has already been constructed, i.e. that the subcategorization frame produced by the syntactic analysis has already been mapped to a corresponding concept in C . This presupposes the existence of a lexicon mapping proteins, protein domains, etc. to the corresponding semantic type, a taxonomy between these types as well as mapping rules from syntactic structures to corresponding concepts. Besides, a syntactic/semantic analysis will be assumed in which a sentence without subject is interpreted as having the protein mentioned in the corresponding *Names*-slot as such. Furthermore, it will be assumed that each *Names*-slot introduces a semantic representation of the protein in question such that it can be referred to. Personal as well as possessive pronouns will be assumed as already resolved. For the sake of simplicity, modal embeddings of verbal phrases such as by the modal *may* will be ignored.

Let's first discuss the example presented in the introduction:

- (63) <Names> U1 small nuclear ribonucleoprotein A (U1 snRNP A protein).
 </Names>
 <FUNCTION> BINDS STEM LOOP II OF U1 SNRNA. [...] < THIS
 INTERACTION> IS REQUIRED FOR THE SUBSEQUENT BINDING
 OF U2 SN-RNP AND THE U4/U6/U5 TRI-SN-RNP. </FUNCTION>

Regarding this example, the key issue is to explain how we get the analysis that the interaction mentioned in the second sentence refers to the binding mentioned in the first one. Only then will we get the overall interpretation that it is this binding which is "REQUIRED FOR THE SUBSEQUENT BINDING OF U2 SN-RNP AND THE U4/U6/U5 TRI-SN-RNP." Let's assume that the binding and the interaction have been mapped to the following DRSs representing concepts in C :



The conceptual representation of the binding can be paraphrased as: “An unknown domain of U1 snRNP A protein binds to the stem loop II domain of U1 snRNA.” The interaction in the second sentence - as there are no arguments given - is mapped to the most general type of biochemical interaction in C , which is the one between two chemical substances.

Provided that we know that a RNA is a nucleic acid and that a nucleic acid and a protein are chemical substances, we can also assume that the following taxonomic relations are in T :

$$\forall e, p, d_1, r, d_2 ((e : bind(p, d_1, r, d_2) \wedge \quad (64)$$

$$protein(p) \wedge protein_domain(d_1) \wedge subunit(d_1, p) \wedge \\ rna_molecule(r) \wedge rna_domain(d_2) \wedge subunit(d_2, r)) \rightarrow \\ (e : bind(p, r) \wedge protein(p) \wedge nucleic_acid(r)) \quad (65)$$

$$\forall e, p, n ((e : bind(p, n) \wedge protein(p) \wedge nucleic_acid(n)) \rightarrow \\ (e : bind(p, n) \wedge protein(p) \wedge chemical_substance(n)) \quad (66)$$

$$\forall e, p, c ((e : bind(p, c) \wedge protein(p) \wedge chemical_substance(c)) \rightarrow \\ (e : bind(p, c) \wedge chemical_substance(p) \wedge chemical_substance(c))) \quad (67)$$

$$\forall e, c_1, c_2 ((e : bind(c_1, c_2) \wedge chemical_substance(c_1) \wedge \\ chemical_substance(c_2)) \rightarrow (e : biochemical_interaction(c_1, c_2) \wedge \\ chemical_substance(c_1) \wedge chemical_substance(c_2))) \quad (68)$$

Because of these taxonomic relations we can conclude that $K_1 \leq_O^* K'_1$, where

K'_1 :	e, p_1, r <hr/> $e : biochemical_interaction(p_1, r)$ $chemical_substance(p_1)$ $chemical_substance(r)$ $p_1 = 'U1\ snRNP\ A\ protein'$ $r = 'snRNA'$
----------	---

According to the definition of m -suitability (9) we can further conclude that K_α is m -suitable to K'_1 so that finally because of definition (18) K_α is interpreted as a bridging reference to K_1 . Thus we yield the following resolution of K_α in which the relation between e and e' has been resolved to identity:

K_α :	e', s_1, s_2 <hr/> $e': \text{biochemical_interaction}(s_1, s_2)$ $\text{chemical_substance}(s_1)$ $\text{chemical_substance}(s_2)$ $s_1 = p_1$ $s_2 = r$ $e' = e$
--------------	---

The following example is somewhat simpler in the sense that it is also resolved to identity but because of *linking* instead of *bridging*:

(69) <Names> Dead ringer like-1 protein (B-cell regulator of IgH transcription) (Bright). </Names>
 <FUNCTION> BINDS A VH PROMOTER PROXIMAL SITE NECESSARY FOR INDUCED MU-HEAVY-CHAIN TRANSCRIPTION. <BINDS> THE MINOR GROOVE OF A RESTRICTED ATC SEQUENCE THAT IS SUFFICIENT FOR NUCLEAR MATRIX ASSOCIATION. [...] </FUNCTION>

The representation of the two binding events in the first and second sentence will be assumed as follows:

K_1 :	e, p_1, d_1, d, d_2, s_1 <hr/> $e: \text{bind}(p_1, d_1, d, d_2, s_1)$ $\text{protein}(p_1)$ $\text{protein_domain}(d_1)$ $\text{subunit}(d_1, p_1)$ $\text{dna_molecule}(d)$ $\text{dna_domain}(d_2)$ $\text{subunit}(d_2, d)$ $\text{dna_sequence}(s_1)$ $p_1 = \text{'Dead ringer like-1 protein'}$ $d_1 = ?$ $d = ?$ $d_2 = \text{'VH promoter proximal site'}$ $s_1 = ?$
$K_\alpha : \alpha :$	$e', p_2, d_3, d', d_4, s_2$ <hr/> $e': \text{bind}(p_2, d_3, d', d_4, s_2)$ $\text{protein}(p_2)$ $\text{protein_domain}(d_2)$ $\text{subunit}(d_2, p_2)$ $\text{dna_molecule}(d')$ $\text{dna_domain}(d_4)$ $\text{subunit}(d_4, d')$ $\text{dna_sequence}(s_2)$ $p_2 = \text{'Dead ringer like-1 protein'}$ $d_3 = ?$ $d' = ?$ $d_4 = ?$ $s_2 = \text{'ATC'}$

The meaning of the DRS K_1 can be paraphrased as "An underspecified domain of the Dead ringer like-1 protein binds a DNA molecule at an underspecified sequence of a VH promoter proximal site". The meaning of the second DRS would then be: "An underspecified domain of the Dead ringer like-1 protein binds a DNA molecule at an ATC sequence of an underspecified DNA region or

domain.”

It is obvious that the DRS K_α is m -suitable to the DRS K_1 . Thus, according to the definition of *linking* (3), we will yield the following resolution of the DRS K_α :

K_α :	$e', p_2, d_3, d', d_4, s_2$ $e': bind(p_2, d_3, d', d_4, s_2)$ $protein(p_2)$ $protein_domain(d_3)$ $subunit(d_3, p_2)$ $dna_molecule(d')$ $dna_domain(d_4)$ $subunit(d_4, d')$ $dna_sequence(s_2)$ $p_2 = 'Dead\ ringer\ like-1\ protein'$ $p_2 = p_1$ $d_3 = d_1$ $d' = d$ $d_4 = d_2$ $s_2 = s_1$ $s_2 = 'ATC'$ $e' = e$
--------------	--

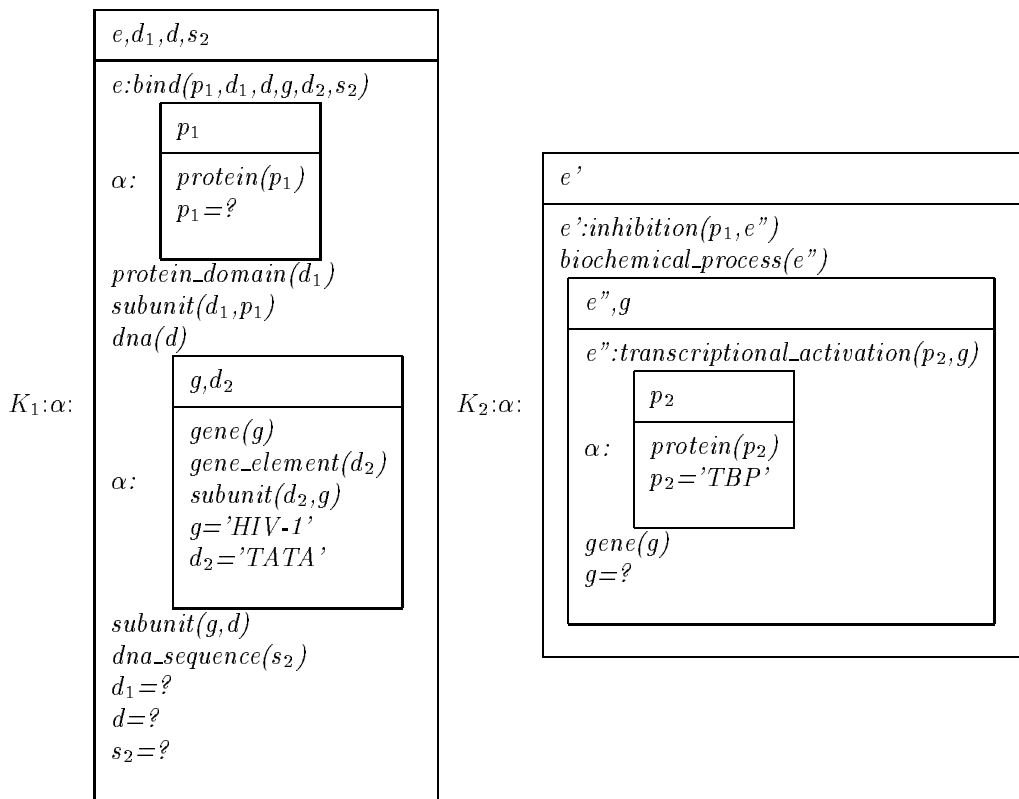
In the following example from the introduction, it is clear that the inhibit-event has to be interpreted as a result of the binding mentioned before. However, both events are connected in a coordinative manner such that the resultative relation is not explicit and thus has to be inferred by the bridging reference resolution process:

(70) <Names> TATA element modulatory factor (TMF). </Names>
 <FUNCTION> THIS PROTEIN BINDS THE HIV-1 TATA
 ELEMENT AND <INHIBITS> TRANSCRIPTIONAL ACTIVATION
 BY THE TATA-BINDING PROTEIN (TBP). </FUNCTION>

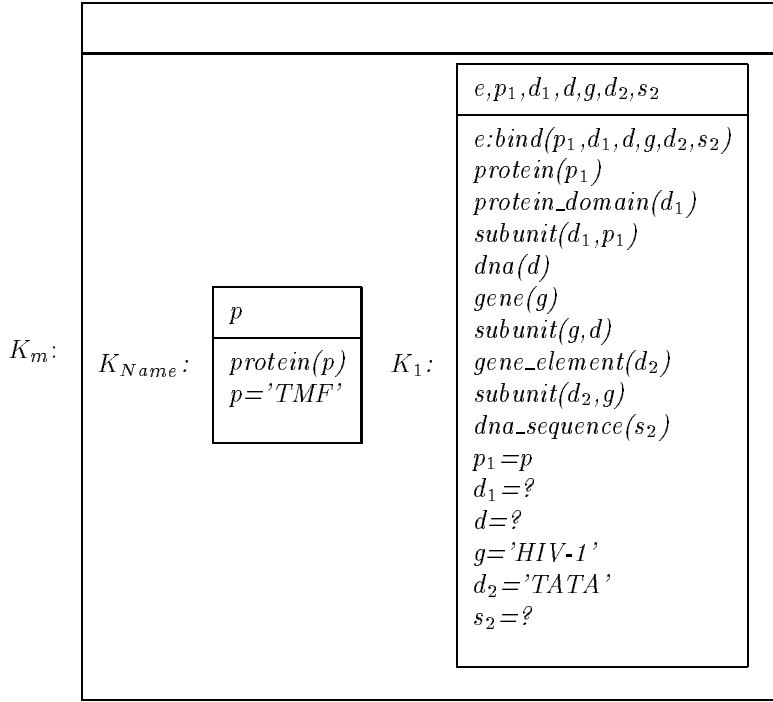
The following DRS will be assumed for the *Names* slot and incorporated into the main DRS K_m :

K_m :	<table border="1" style="margin-left: 20px;"> <tr> <td style="padding-right: 10px;">K_{Name}:</td> <td style="border: 1px solid black; padding: 5px;"> p $protein(p)$ $p = 'TMF'$ </td> </tr> </table>	K_{Name} :	p $protein(p)$ $p = 'TMF'$
K_{Name} :	p $protein(p)$ $p = 'TMF'$		

The binding and inhibition events will be represented respectively by the α -marked DRSs K_1 and K_2 :



Let's now discuss the resolution of this example step by step. The definite descriptions 'THIS PROTEIN' and 'THE HIV-1 TATA ELEMENT' in K_1 as well as the definite description 'THE TATA-BINDING PROTEIN (TBP)' in K_2 are all α -marked. In fact, this is in line with the widely agreed analysis of definite descriptions as presupposition triggers. So, before resolving the relation between the binding event and some previous event thus establishing discourse coherence, first the definite descriptions 'THIS PROTEIN' and 'THE HIV-1 TATA ELEMENT' have to be resolved. The author will gloss over the details of this resolution, but ask the reader to verify that p_1 is identified with p via *linking* and the second α -marked DRS is locally accommodated as there is no suitable antecedent. In addition, the binding event is also accommodated as there is no previous event it could be related to. So the DRS K_1 is (after having been merged with the main DRS K_m) resolved to:

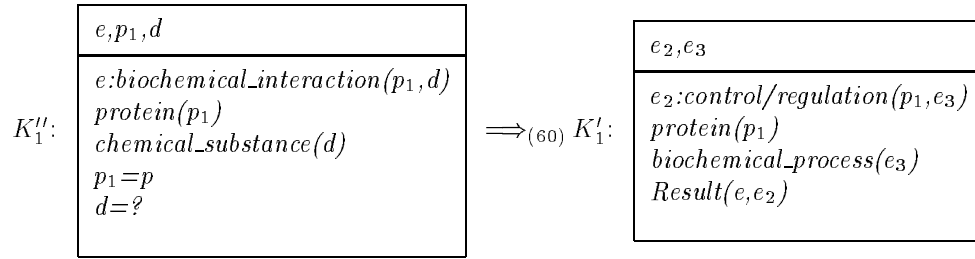


Before resolving the DRS K_2 , the embedded α -marked DRS has to be resolved first. Actually, p_3 has to be locally accommodated as there is no suitable antecedent representing a protein with value 'TBP'.

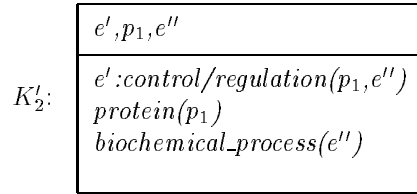
If we assume the taxonomic relations (71) and (72) in T, together with the relations (64) - (68) it will be the case that $K_1 \leq_O^* K_1''$ and because of the conceptual definition that a biochemical interaction normally results in the regulation/control of a certain biochemical process (60) that $K_1'' \implies_{60} K_1'$ (see below) and consequently that $K_1 \implies_O K_1'$ by definition 17.

$$\begin{aligned}
& \forall e, p, d_1, d, g, d_2, s ((e : bind(p, d_1, d, g, d_2, s) \wedge protein(p) \wedge \\
& protein_domain(d_1) \wedge subunit(d_1, p) \wedge dna(d) \wedge gene(g) \wedge \\
& subunit(g, d) \wedge gene_element(d_2) \wedge subunit(d_2, g) \wedge dna_sequence(s)) \rightarrow \\
& (e : bind(p, d_1, d, g, s) \wedge protein(p) \wedge protein_domain(d_1) \wedge subunit(d_1, p) \\
& dna(d) \wedge dna_domain(g) \wedge subunit(g, d) \wedge dna_sequence(s))) \quad (71)
\end{aligned}$$

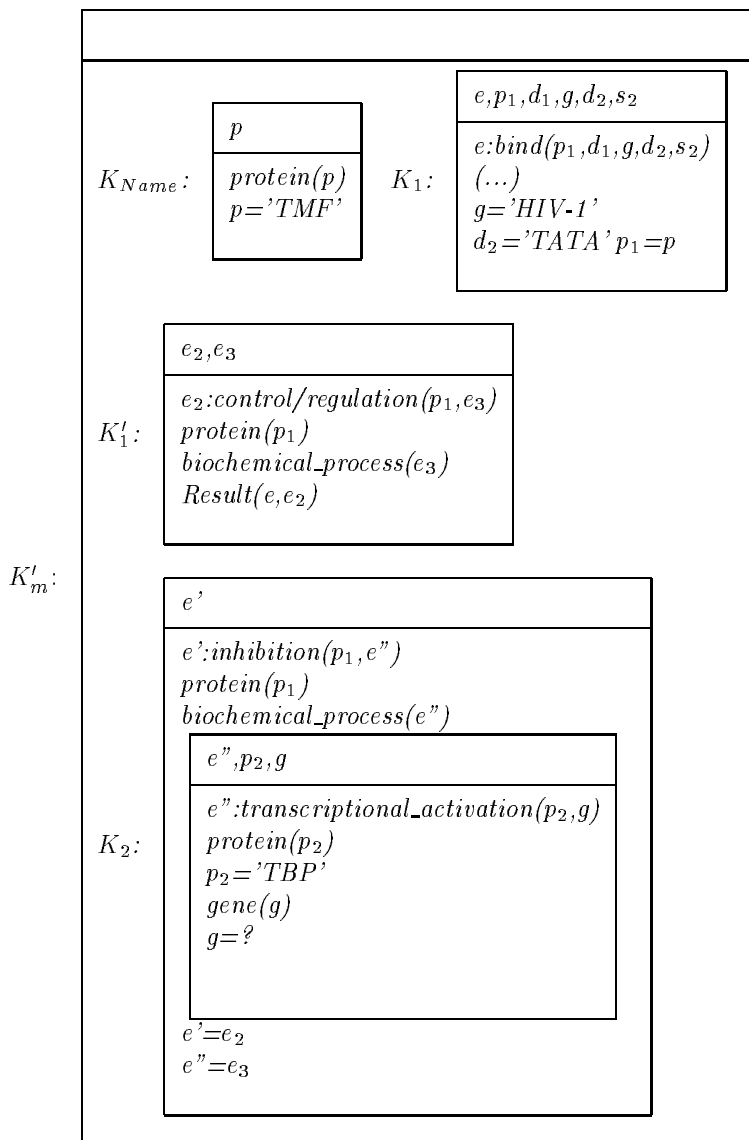
$$\begin{aligned}
& \forall e, p, d_1, d, g, s ((e : bind(p, d_1, d_2, g, s) \wedge protein(p) \wedge \\
& protein_domain(d_1) \wedge \\
& subunit(d_1, p) \wedge dna(d) \wedge \\
& dna_domain(d_2) \wedge subunit(d_2, d) \wedge dna_sequence(s)) \rightarrow \\
& e : bind(p, d) \wedge protein(p) \wedge nucleic_acid(d)) \quad (72)
\end{aligned}$$



On the other hand, because of the taxonomic relation (55) it is also the case that $K_2 \leq_O^* K_2'$, where



Finally, as K_2' is m -suitable to K_1' , according to definition (18) K_1' will be accommodated and K_2 interpreted as referring to K_1 thus yielding the following resolution K_m' :



Thus because $Result(e, e_2)$ and $e' = e_2$, we finally get $Result(e, e')$ as the relation between the binding e and the inhibit event e' , which is exactly what we wanted.

The relation between the repression and the binding in the next example is an explanatory one, i.e. the binding explains the fact that the repression occurs:

(73) <Names> CreA protein. </Names>
 <FUNCTION> [...] REPRESSES THE TRANSCRIPTION OF THE
 ALCR, ALCA AND ALDA GENES. <BINDS> TO A GC-RICH
 REGION IN THEIR PROMOTER. </FUNCTION>

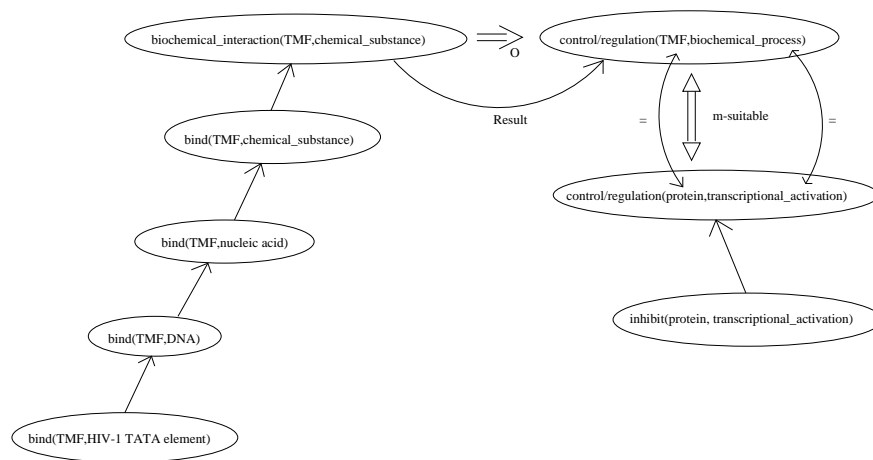
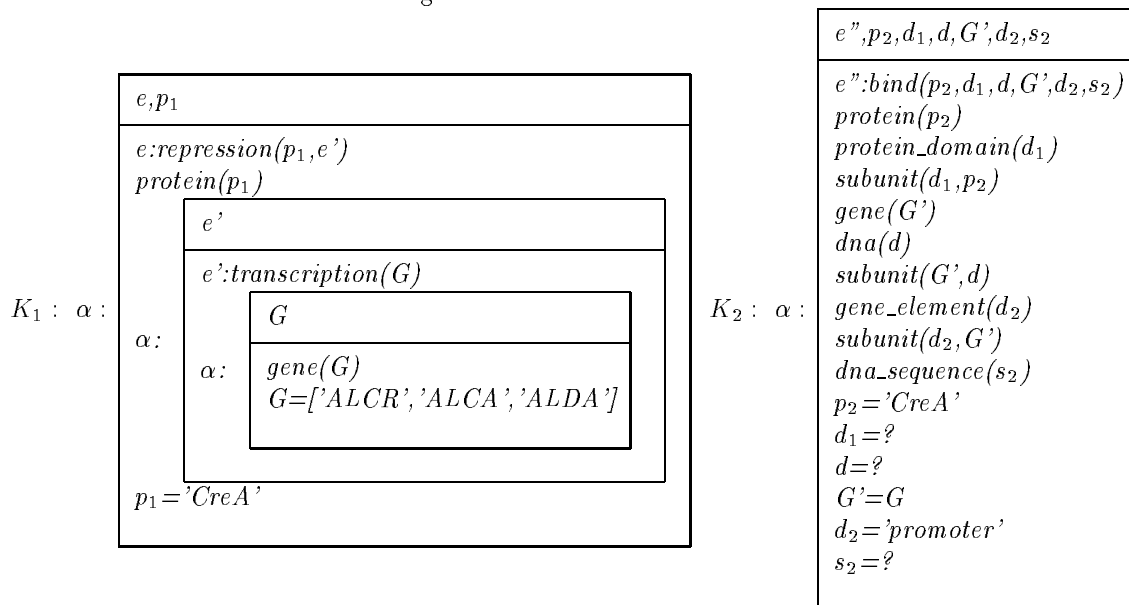


Figure 4: Schematic visualization of the resolution of example (69)

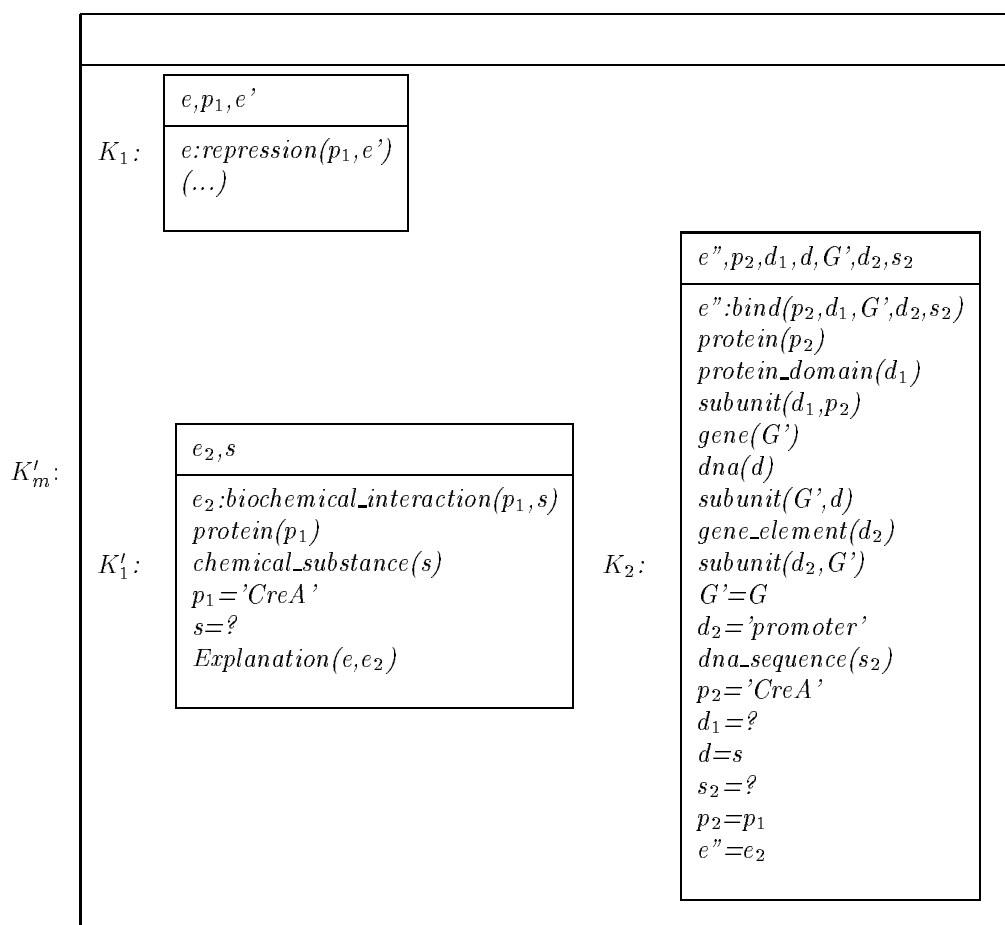
The relevant DRSs are the following ones:



15

First of all the definite description 'THE ALCR, ALCA AND ALDA GENES' has to be resolved as it is the most deeply embedded referring expression. But

¹⁵Here the subject has been analyzed as being the protein in the *Names*-slot.



So finally because $Explanation(e, e_2)$ and $e'' = e_2$ we get $Explanation(e, e'')$.

The following example will be resolved analogously to the previous example so that the *inhibit*-event will be interpreted as a result of the *binding* mentioned before:

(74) <Names> Ornithine decarboxylase antizyme (ODC-Az). </Names>
<FUNCTION> BINDS TO, AND DESTABILIZES, ORNITHINE
DECARBOXYLASE WHICH IS THEN DEGRADED. ALSO INHIBITS
CELLULAR UPTAKE OF POLYAMINES BY INACTIVATING THE
POLYAMINE UPTAKE TRANSPORTER. </FUNCTION>

However, as personal communication with biologists revealed, this interpretation is definitely wrong. So it becomes clear that the approach also leads in some cases to erroneous interpretations which are due to the generalizing form certain conceptual definitions are specified within the ontology.

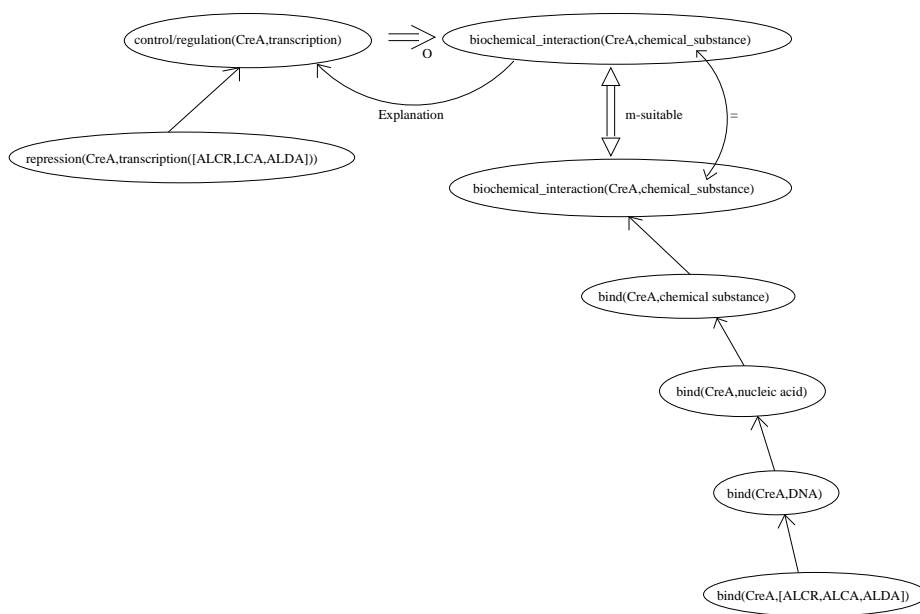


Figure 5: Schematic visualization of the resolution of example (73)

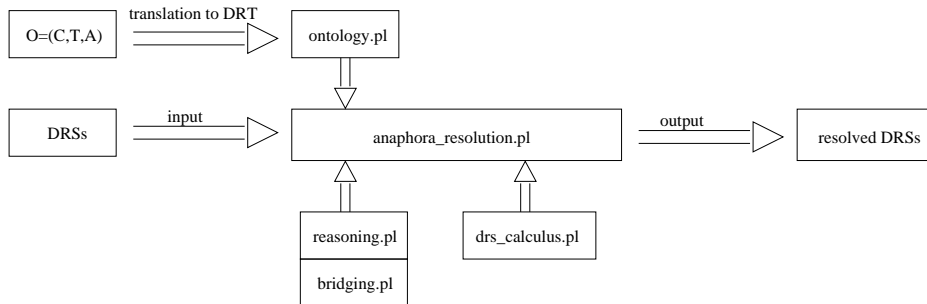


Figure 6: System architecture

7 Implementation

The system has been implemented in Prolog and is based on the projection algorithm of Blackburn and Bos ([4]). During the design of the system special attention has been paid to its modularity and reusability. Figure 6 shows the different modules that constitute the whole system. The module *anaphora.pl* represents the core of the system. There the projection and *m*-suitability algorithms are defined. The ontology is coded in the *ontology.pl* module and can be exchanged depending on the domain of application. The *reasoning.pl* module implements the inference mechanism described in section 4.2, while *bridging.pl* contains the ontology-based definitions of the *linking*, *bridging* and *accommodation* operators (compare section 4). The last two modules would have to be exchanged if some knowledge representation formalism other than an ontology should be used. Finally, the module *drs_calculus.pl* implements the Generalized Modus Ponens and could be extended to contain other inference rules defined in [27]. In the following, these five modules are described in more detail.

7.1 anaphora_resolution.pl

As already mentioned and visualized in figure 6, the *anaphora_resolution.pl* module represents the invariant core of the system. It contains a slightly modified version of Bos et al.'s presupposition projection algorithm ([4]). Here it is:

```
projectDRS([merge(D1,D2)],K,Res) :- projectDRS([D1],K,K2),
                                   projectDRS([D2],K2,Res).

projectDRS([alfa(drs(D,C))],K,Res) :- projectConds(C,K,K2,C2,R1),
                                       ar(drs(D,C2),K2,K3,R2),
                                       append(C2,R1,C3),
                                       append(C3,R2,C4),
```

```
append([drs(D,C4)],K3,Res).
```

```
projectDRS([drs(D,C)],K,Res) :- projectConds(C,K,K2,C2,B),
                                append(C,C2,C3),
                                append(C3,B,C4),
                                append([drs(D,C4)],K2,Res).
```

The abstract signature of the `projectDRS` predicate is:

```
projectDRS(+D,+K,-R)
```

Basically it takes a DRS D and resolves it with regard to the context K thus yielding a resolved DRS R . When projecting an α -marked DRS the anaphora resolution algorithm (*ar*) will be called. As already mentioned in section 4, anaphora resolution is a recursive process in the sense that the most deeply embedded anaphor has to be resolved first. However, as the above definitions show, the *projectDRS* predicate is not recursive. In fact, the *projectDRS* predicate is only called at the top level; it is the *projectConds* predicate which is responsible for the recursion:

```
projectConds([],K,K,[],[]).
```

```
projectConds([alfa(drs(D,C))|R1],K,K4,Res,B5) :-
    projectConds(C,K,K2,R,B1),
    ar(drs(D,R),K2,K3,B2),
    append(B1,B2,B3),
    projectConds(R1,K3,K4,R3,B4),
    append(R,R3,Res),
    append(B3,B4,B5), !.
```

```
projectConds([or(D1,D2)|R1],K,K4,R3,B5) :-
    projectConds([D1],K,K2,D3,B1),
    projectConds([D2],K2,K3,D4,B2),
    append(B1,B2,B3),
    projectConds(R1,K3,K4,R2,B4),
    append([or(D3,D4)],R2,R3),
    append(B3,B4,B5), !.
```

```
projectConds([not(D1)|R1],K,K2,[not(D1)|R2],B) :-
    projectConds(R1,K,K2,R2,B).
```

```
projectConds([drs(D,C)|R1],K,K3,[drs(D,C2)|R2],B3) :-
    projectConds(C,K,K2,C2,B1),
    projectConds(R1,K2,K3,R2,B2),
    append(B1,B2,B3), !.
```

`projectConds([X|R1],K,K2,[X|R2],B) :- projectConds(R1,K,K2,R2,B).`

The abstract signature of the *projectConds* predicate is as follows:

`projectConds(+D,+K,-K2,-R,-B)`

where D is the DRS to be projected, K the corresponding context, K2 the updated context - i.e. containing the information accommodated by the anaphora resolution process -, R the DRS D without α -marks and B the set of discourse referent bindings of the form $d_1 = d_2$ yielded as result of the anaphora resolution. It should be noted that at the top level of the recursion the result R of the *projectDRS* predicate is exactly the union of the updated context K2, the DRS R and the set of bindings B as returned by *projectConds*. Certainly then the question arises why the updated context, the DRS R and the set of bindings are kept within *projectConds* and returned to the previous level of recursion separately. The reason is that due to performance reasons the values of discourse referents have been implemented as Prolog constants and not as strings, i.e. `p='JAK'` instead of `p="JAK"`. This in turn leads to the fact that the system can not distinguish between a value assignment such as `p='Jak'` and the identification of two discourse referents `p=p'` as yielded by the anaphora resolution. Thus the accommodation of such an identification of discourse referents at a level other than the top one can hinder the whole process from being successful because the introduction of equations as above are interpreted as additional value assignments which affect the suitability of the involved DRSs. Several solutions to this problem are possible. Certainly, the most straightforward one is to represent values of discourse referents as strings instead of constants. However, this has turned out to significantly increase the processing time of the system. On the other hand, it could be thought of introducing a special 'value'-predicate for the assignment of a value to a discourse referent, i.e. `value(p,'JAK')` or a special 'id'-predicate for the identification of two discourse referents, i.e. `id(p,p')`. However, both solutions seem rather inelegant. The author has thus opted for the solution mentioned first which seems the most straightforward and intuitive, i.e. the accommodation of the bindings only at the top level of projection.

Besides establishing the recursion to resolve the most deeply embedded anaphor first, the *projectConds* clause is also responsible for specifying special projection rules for disjunctive ($K_1 \vee K_2$) and negated conditions ($\neg K_1$). Conditionals of the form $K_1 \implies K_2$ have not been considered in the implementation above as they turned out to be very rare in the corpus, but it should actually represent no difficulty to extend the *projectConds* predicate to cope with them.

The *anaphora_resolution* module also specifies the way anaphora resolution is made determinate and in particular the application order of the three operations *linking*, *bridging* and *accommodation*:

`ar(A,K,K2,B) :- linking(A,K,K2,B) ,!`
`ar(A,K,K2,B) :- bridging(A,K,K2,B) ,!`
`ar(A,[D|K1],[D|K2],B) :- ar(A,K1,K2,B) ,!`

```
ar(A,K,K2,B) :- accommodate(A,K,K2,B).
```

The above implementation expresses the preference for *linking to bridging* and presents *accommodation* as the last resource for when no suitable antecedent for A can be found within the context K. The abstract signature of the *ar* predicate is as follows:

```
ar(+A, +K, -K2, -B)
```

where A is the DRS to be resolved with regard to the context K, K2 is the updated context and B is the set of discourse referent identifications yielded by the resolution.

The *anaphora_resolution* module also contains the implementation of the notion of *m-suitability*. However, the details concerning its implementation will be omitted here as it is basically a one to one translation of definition (9) to Prolog and in addition it is not crucial for the understanding of the system as a whole.¹⁶

7.2 ontology.pl

As already mentioned, the module *ontology.pl* encodes the ontology. As figure 6 shows, this module basically represents the product of the translation from the knowledge representation language used in the ontology to DRT. However, the taxonomic relations in T have to be logically separated from the conceptual definitions in D. For this purpose the predicates *is_a* and *def* have been used. The translation of the ontology representing world knowledge about rooms (see section 4) for example would look as follows:

```
is_a(drs([x],[chandelier(x)]),drs([],[lamp(x)])).
def(drs([x],[room(x)]),drs([y],[lamp(y),part_of(y,x)])).
```

7.3 bridging.pl

The *bridging.pl* module basically implements the definitions of the *linking*, *bridging* and *accommodation*-operators as defined in section 4:

```
linking(drs(_,C1),[drs(D2,C2)|K], [drs(D2,C2)|K],M) :-
    m_suitable(C1,C2,[],M), !.
```

```
bridging(drs(D1,C1),[drs(D2,C2)|K],[drs(D2,C2)|K],M) :-
    is_a*(drs(D1,C1),drs(_,C3)),
    m_suitable(C3,C2,[],M), !.
```

```
bridging(drs(_,C1),[drs(D2,C2)|K],[drs(D2,C)|K],M) :-
    imp_o(drs(D2,C2),drs(D3,C3)),
```

¹⁶See the appendix for the full source code.

```

append(C2,[drs(D3,C3)],C),
m_suitable(C1,C3,[],M),!.

```

```

bridging(drs(D1,C1),[drs(D2,C2)|K],[drs(D2,C)|K],M):-
  is_a*(drs(D1,C1),drs(_,C4)),
  imp_o(drs(D2,C2),drs(D6,C6)),
  m_suitable(C4,C6,[],M),
  append(C2,[drs(D6,C6)],C),!.

```

```

accommodate(drs(_,_),K,K,[]).

```

The first *bridging* clause corresponds to the first part of the *bridging* definition (18), while the second and third clauses correspond to the second part of it. The *is_a** and *imp_o* predicates respectively represent the transitive closure \leq_o^* of $<_o$ and the implication with regard to an ontology \implies_o (17) as defined in section 4.2.

7.4 reasoning.pl

In the module *reasoning.pl* the ontology-based inference mechanism as defined in section 4.2 has been implemented. In order to 'hide' the implementation details to the rest of the application, the *is_a** and *imp_o* predicates have been merely defined as an interface:

```

is_a*(drs(D1,C1), drs(D2,C2)) :- isra(drs(D1,C3),drs(D2,C2)),
                                match(C3,C1).

```

```

imp_o(drs(D1,C1),drs(D2,C2)) :- imp(drs(D1,C3),drs(D2,C2)),
                                match(C3,C1).

```

The *match* predicate basically guarantees that the success of the *is_a** and *imp_o* predicates is not dependent on the order of the conditions in C1:

```

match([],-).
match([E1|R1],R) :- member(E1,R,R2), match(R1,R2).

```

The implementation of the predicates *isra* and *imp* is now straightforward as it corresponds one to one to the recursive definitions of specialization between DRSs (16) and implication with regard to an ontology defined in section 4.2:

```

isra(DRS1,DRS2) :- is_a(DRS1,DRS2).

```

```

isra(drs(D1,C1),drs(D3,C3)) :- is_a(drs(D1,C1),drs(_,C2)),
                                isra(drs(_,C2),drs(D3,C3)).

```

```

imp(DRS1,DRS2) :- imply(DRS1,DRS2).

```

```

imp(DRS1,DRS2) :- imply(drs(_,C3),DRS2),
                  isra(DRS1,drs(_,C3)).

```

However, a few words should be said about the implementation of the end conditions *is_a* and *imply* of the above recursive definitions which respectively correspond to definition (16) and the end condition of the recursive definition (17). Prolog provides a nice way to implement these end conditions by making use of its unification mechanism. Thus, instead of directly implementing the implication between DRSs $K_1 \implies_{[K \implies K']} K_2$ by defining a homomorphism such as in (9) and a copy mechanism such as in (10), the fact that the DRS K_1 matches the DRS K will be ensured through unification of their conditions and discourse referents represented by Prolog-variables. Thus, the Prolog variables in K' will get automatically instantiated with the correct values so that we will yield an alphabetic variant of K' through unification. The only thing left then is to instantiate the free variables occurring in K' - corresponding to discourse referents declared in K' - with new discourse referents. The implementation of the end condition *imply* is for example as follows:

```
imply(DRS1,DRS2) :- def(DRS1,DRS2), gmp(DRS1,DRS2).
```

The above definition basically states that the DRS1 implies the DRS2 if there is a corresponding conceptual definition in the ontology which can be respectively unified with DRS1 and DRS2. The predicate *gmp* has merely the function of instantiating the Prolog-variables representing discourse referents declared within DRS2 with new values. This predicate is described in the next section.

7.5 drs_calculus.pl

The module *drs_calculus.pl* is intended to be a unification-based implementation of the DRS calculus defined by Kamp et Reyle. However, for the purposes of this diploma thesis it has sufficed to implement the Generalized Modus Ponens by the following clause:

```
gmp(drs(-,-),drs(D2,-)) :- new(D2).
```

In fact, as mentioned in the previous section, the unification-based implementation of the GMP boils down to instantiating the Prolog variables representing discourse referents in D2 with new constants. The implementation of the *new*-predicate is not given here as the mechanism of how to (globally) store and increase a certain index and thus create a new constant is too dependent on the Prolog version used.

8 Evaluation and Results

8.1 Training and Testing

In order to verify the utility and scalability of the approach presented within this work, a quantitative evaluation of its performance has been carried out. Typically within computational linguistics research and in particular in the field of information extraction ([37]), such an evaluation of the performance of an approach involves the development of it on training data and the subsequent verification of its scalability on unseen or test data.

For this purpose the corpus described in section 2 has been divided into a training and a test corpus each consisting of 250 entries. In both the training and test corpus verbs and definite descriptions representing events, states or entities have been marked by the author and assigned a unique identifier. The criteria for identifying a verb or definite description as an event have been outlined already in section 5.2. Events have been marked with the label 'e' followed by a unique numeral ID with the only exception of binding events which have been marked with the label 'b'. States and entities have been respectively marked by an 's' and a 'd'. Here's an example taken from the training corpus:

```
(75) <Names> Acyl-CoA-binding protein homolog (ACBP) (Diazepam
binding inhibitor homolog) (DBI). </Names>
<FUNCTION> MAY <id="s17" PLAY> IMPORTANT FUNCTIONS
IN <id="e766" THE CONTROL OF BRAIN AND PITUITARY
ACTIVITIES>. MAY <id="e18" REGULATE> GABA
NEUROTRANSMISSION THROUGH A PARACRINE AND/OR
AUTOCRINE MECHANISM. MAY NOT <id="b2" BIND>
ACYL-COA ESTERS. </FUNCTION>
```

Table 4 gives some statistics about the training and test corpora. In particular it indicates the number of tokens, the number of events, states and entities marked as well as the number of definite descriptions of each corpus. The diagram in

	training corpus	test corpus
#tokens	12666	12180
#events	708 (54.05%)	894 (56.69%)
#states	175 (13.36%)	209 (13.25%)
#entities	427 (32.60%)	474 (30.06%)
Total	1310 (100%)	1577 (100%)
#DDs	510	530

Table 4: Statistics of the training and test corpora

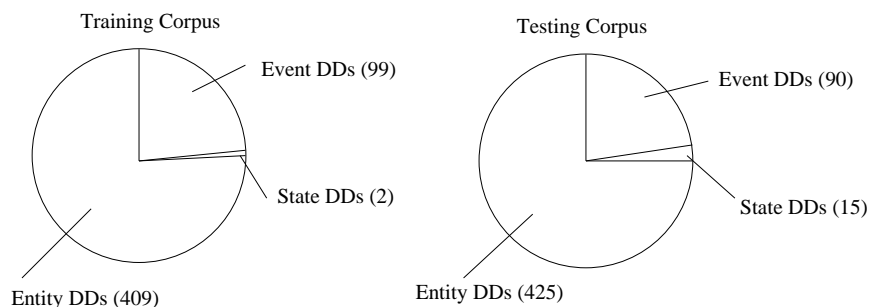


Figure 7: Distribution of the DDs after types

figure 7 shows the distribution of DDs of the training and test corpora after types.

8.2 Ontology Development

The ontology $O_{Bio} = (C_{Bio}, T_{Bio}, D_{Bio})$ has been developed on the textual basis of the training corpus in the sense that:

1. Suitable conceptual representations of events, states and entities have been developed thus constituting the set C_{Bio} .
2. The taxonomic relations between these concepts necessary for the bridging reference resolution process to be successful have been expressed through logical axioms forming the set T_{Bio} .
3. The conceptual relations between different events or states have also been captured in form of logical axioms. These axioms form the set D_{Bio} consisting of partial definitions of concepts.

Within this ontology development step, special attention has been paid to represent only those concepts as well as those taxonomic and conceptual relations having a certain degree of relevance and generality. The aim has been to yield an ontology which is not too specifically tailored to the corpus it was developed on thus being potentially reusable across different biochemical texts. Table 5 summarizes some facts about the ontology. After developing the ontology, the events, states and entities of both the training and test corpus have been manually mapped to DRSs representing the corresponding ontological concept in C_{Bio} . In order to verify the reusability and generality of the created ontology, the ontological coverage of events, states and entities for each corpus has been determined. Table 6 shows that there is a significant loss of ontological coverage when comparing the test corpus to the training corpus. This loss in ontological

number of concepts ($ C_{Bio} $)	129
number of taxonomic relations ($ T_{Bio} $)	50
number of axiomatic definitions ($ D_{Bio} $)	19

Table 5: Some facts about O_{Bio}

	training corpus	test corpus
Covered Events	46.49%	37.92%
Covered States	1.69%	3.83%
Covered Entities	28.47%	18.78%
Total Coverage	34.63%	27.65%

Table 6: Ontological coverage of the training and test corpora

coverage is clearly due to the fact that the ontology has been developed on the training corpus. However, it is not clear how this loss can be quantitatively interpreted in terms of the generality, scalability and reusability of the developed ontology. Furthermore, the above results only concern the generality of the concepts in C_{Bio} . In order to evaluate the generality of the taxonomic relations in T_{Bio} as well as the partial definitions of D_{Bio} , the results of the bridging resolution process on the training and test corpora have to be compared.

8.3 Agreement between Annotators

In order to evaluate the performance of the bridging resolution approach in a quantitative manner, the training and test corpora have been annotated by different subjects with the expected discourse relations by making use of the MMAX annotation tool developed by Müller et al. ([42]). The relevant events, states and entities had been previously marked by the author so that the task of the annotators has basically been to choose the appropriate relation between two marked expressions out of table 7. The Explanation/Elaboration relation is defined as the disjunction of the Explanation and Elaboration relations considered by Lascarides et al. ([35]). The reason why they have been collapsed into one relation is that the distinction between them has been expected to be difficult for the annotators. The training corpus has been annotated only by the author, while the test corpus has been annotated independently from each other by the author and two biologists. The agreement between the annotators has been measured with the kappa statistic ([9]). The kappa coefficient (K) measures the pairwise agreement among a set of individuals making category

	event
event	Coreference Result Elaboration/Explanation
state	Result Elaboration/Explanation
entity	Role

Table 7: Discourse relations considered for the annotation

K	Agreement
< 0.00	poor
0.00 - 0.20	slight
0.21 - 0.40	fair
0.41 - 0.60	moderate
0.61 - 0.80	substantial
0.81 - 1.00	almost perfect

Table 8: Classification by Landis and Koch of the agreement as measured by the kappa statistic

judgments by taking into account the expected agreement by chance:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad (76)$$

where $P(A)$ is the proportion of times the annotators agree and $P(E)$ is the proportion of times they would be expected to agree by chance.

Due to performance reasons concerning the annotation tool, the test corpus has been divided into 25 blocks each consisting of 10 SWISS-PROT entries. The graph in figure 8 shows the kappa coefficient for each of the 25 blocks as well as the overall agreement which is $K=0.31$.¹⁷ Following the classification by Landis and Koch ([33]) of the agreement as measured by the kappa statistic, this value can be classified as corresponding to a 'fair' agreement between the annotators (compare table 8). Certainly, the agreement is not an overwhelming one, but on the other hand it definitely shows that annotators agree to a certain degree in determining the rhetorical relations between events as antecedents and other referring expressions. Furthermore, it shows that the task of determining these relations is not a trivial one and that it is quite subjective. This observation

¹⁷The first block has been annotated by the two biologists together as a first test to get familiar with the tool and the annotation. Without considering this block, the overall agreement amounts to $K=0.29$.

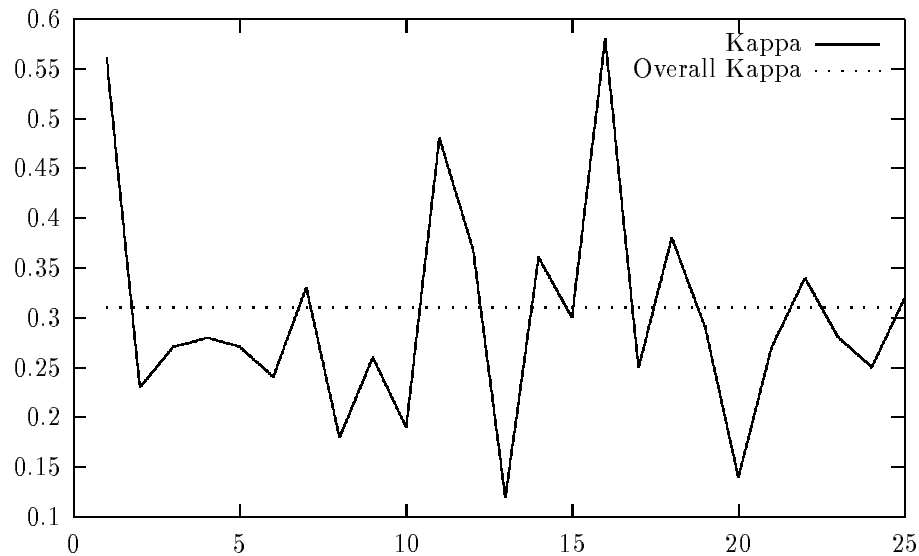


Figure 8: Agreement of the annotators over the 25 blocks

already points to the limits of a machine-based approach. Certainly, a machine-based approach will not be expected to outperform a human subject on such a highly subjective task. Thus, by analogy with the famous Turing-test ([68]), if the output of such an approach can not be identified as being non-human, this has to be regarded as a definite success.

8.4 Results

The performance of the approach outlined in section 4 on the training and test corpus has been measured in terms of precision and recall against a certain standard. The approach described in section 4 yielded a recall of $R=52.57\%$ and a precision of $P=84.40\%$ and thus $F=64.79\%$ measured against the author's annotation of the training corpus. Table 9 shows the precision and recall for each type of relation. The performance of the approach on the test corpus has been measured against the following three standards:

- **AUTHOR**: the set of discourse relations annotated by the author
- **2/3**: the set of discourse relations on which at least two of the three annotators agree
- **ALL**: the set of discourse relations on which all three annotators agree

	Recall	Precision
Coreference	80%	100%
Result	48.88%	81.48%
Explanation/Elaboration	40.35%	88.46%
Role	100%	76.47%

Table 9: Precision and Recall on the training corpus for each type of relation

	Recall	Precision	F-measure
AUTHOR	53.84%	81.15%	64.73%
2/3	38.31%	83.61%	52.54%
ALL	54.54%	83.61%	66.01%

Table 10: Results of the bridging reference resolution approach measured against the three 'gold standards'

Table 10 contains the results of the evaluation against the three standards. In the cases of the '2/3' and the 'ALL' standard, precision has been measured against the union of the discourse relations of the three annotators. Otherwise the approach would have been penalized for finding answers which have been actually given by some annotator. The results show that measured against the 'AUTHOR' or the 'ALL' standards, the approach as well as the developed ontology can be regarded as scalable as the results on the training and test corpus are similar from a quantitative point of view. However, it remains to be clarified why the performance measured against the '2/3' standard is significantly worse. This question is (among others) discussed in the next section.

	Recall	Precision
Coreference	55%	84.61%
Result	48.93%	76.67%
Explanation/Elaboration	52.94%	87.10%
Role	78.44%	83.33%

Table 11: Precision and Recall on the test corpus for each type of relation measured against the 'AUTHOR' standard

8.5 Discussion of the Results

8.5.1 Annotation Errors

As already mentioned in section 8.4, the recall on the test corpus measured against the '2/3' standard is significantly worse than measured on the 'AUTHOR' or the 'ALL' standard. In fact, a closer look at the '2/3' standard reveals that many of the discourse relations found by the two biologists are incorrect from a linguistic point of view. Let's consider the following example from the test corpus:

- (77) <Names> Guanine nucleotide-binding protein beta subunit. </Names>
<FUNCTION> GUANINE NUCLEOTIDE-BINDING PROTEINS (G PROTEINS) ARE <id="s68" INVOLVED> AS A MODULATOR OR TRANSDUCER IN VARIOUS TRANSMEMBRANE SIGNALING SYSTEMS. <id="d135" THE BETA AND GAMMA CHAINS> ARE <id="s69" REQUIRED> FOR <id="e187" THE GTPASE ACTIVITY>, FOR <id="e188" REPLACEMENT> OF GDP BY GTP, AND FOR G PROTEIN- EFFECTOR <id="e189" INTERACTION>. <id="s70" REQUIRED> FOR ADENYLATE CYCLASE <id="e190" ACTIVATION>. </FUNCTION>

The two biologists identified the relation Explanation/Elaboration(s70,e190), i.e. the activation of adenylate cyclase explains or elaborates the fact that the G-protein is required for this activation. This interpretation is neither correct from a semantic nor a syntactic point of view. In fact, from a syntactic point of view the PP *FOR ADENYLATE CYCLASE ACTIVATION* is merely a prepositional complement of the verb *REQUIRED* and therefore can not be interpreted as an explanation or elaboration of the state represented by it.

Actually there are 24 such incorrect annotations within the '2/3' standard. In fact, if all these incorrect annotations are eliminated from it, recall raises to R=45.38%. The obvious conclusion to be drawn is that because of these annotation errors the '2/3' standard as such is not suitable for evaluation purposes.

8.5.2 Underspecification

Poesio and Reyle ([50]) suggest that there are cases in which readers/listeners don't appear to have any problems in resolving ambiguous pronouns. These cases are characterized by the fact that the alternative interpretations are part of an underlying mereological structure and that these alternative interpretations are equivalent in a certain sense. A similar phenomenon has been observed in the training and test corpora. Take the following example:

- (78) MAY <id="b37" BIND> TO AND <id="e87" INHIBIT> CYCLIN-DEPENDENT KINASE <id="e88" ACTIVITY>, <id="e89" PREVENTING> <id="e90" PHOSPHORYLATION> OF CRITICAL CYCLIN-DEPENDENT KINASE SUBSTRATES AND <id="e91" BLOCKING> CELL CYCLE PROGRESSION.

It is quite obvious that the *inhibit*-event “e87” has to be interpreted as being resultative of the *binding*-event “b37”, i.e. $\text{Result}(\text{b37}, \text{e87})$ holds. However, the *prevent*-event “e89” can be interpreted as resultative to either the *binding*-event “b37” or the *inhibit*-event “e87”. In the following it will be argued that both interpretations are valid and furthermore equivalent. If the *prevent*-event “e89” is interpreted as a result of the *inhibit*-event “e87”, i.e. $\text{Result}(\text{e87}, \text{e89})$ holds, because of the transitivity of the *Result*-relation we will get that $\text{Result}(\text{b37}, \text{e87}) \wedge \text{Result}(\text{e87}, \text{e89}) \implies \text{Result}(\text{b37}, \text{e89})$, which corresponds exactly to the interpretation of the *prevent*-event “e89” as being resultative of the *binding*-event “b37”. In fact, the annotations of the different subjects show that both interpretations are valid. Thus, as in Poesio and Reyle’s argumentation, we yield that the alternative interpretations are valid and furthermore equivalent. However, in the case described here this is not due to an underlying mereological structure but to the transitivity of the *Result*-relation. The same argumentation holds for the *blocking* event “e91” which, for the above reasons, can be interpreted as resultative of the *prevent*-event “e89”, the *inhibit*-event “e87” or the *binding*-event “b37”.

Appendix A contains further examples from the test corpus where annotators agreed on the (rhetorical) relation but differed in the choice of the antecedent. It is interesting to observe that in many cases the alternative interpretations can be explained in terms of the transitivity argument outlined above.

The fact that the antecedent for some bridging references is ambiguous has consequences for the evaluation of the approach. Thus, if the system identifies one of the possible and correct antecedents, the answer has to be regarded as correct. In fact, in the case of the above example the system calculates the relation $\text{Result}(\text{b37}, \text{e91})$ while, according to the ‘ALL’ standard, $\text{Result}(\text{e89}, \text{e91})$ is the correct answer. Due to the above arguments the system’s answer has to be regarded as correct so that precision increases to $P=84.42\%$.

8.5.3 Recall and Precision

The results in section 8.4 show that almost 50% of the bridging references involving events as antecedents are not resolved. A detailed analysis of the results reveals that in most cases this is due to lack of the necessary world knowledge. Certainly, the ontology could be refined, but there is also a clear trade-off between such a further refinement on the one hand and its generality on the other. In fact, if the ontology would be further developed on the training corpus, this might yield better results on the training corpus itself, but it is not clear if this further refinement would scale up and also produce better results on the test corpus.

One example taken from the test corpus should suffice to show that more detailed and complex domain knowledge is indeed necessary to resolve certain bridging references:

- (79) <Names> Annexin I, isoform P35 (Lipocortin I) (Calpactin II)
(Chromobindin 9) (Phospholipase A2 inhibitory protein). </Names>

<FUNCTION> <id="b8" CALCIUM/PHOSPHOLIPID-BINDING>
 PROTEIN WHICH <id="e16" PROMOTES> MEMBRANE FUSION
 AND IS <id="s4" INVOLVED> IN EXOCYTOSIS. <id="d11" THIS
 PROTEIN> <id="e17" REGULATES> PHOSPHOLIPASE A2
 <id="e18" ACTIVITY>. IT SEEMS TO <id="b9" BIND> FROM
 TWO TO FOUR CALCIUM IONS WITH HIGH AFFINITY.
 </FUNCTION>

In order to get the resultative relation between the state of being 'INVOLVED in EXOCYTOSIS' and the fact that the protein in question 'PROMOTES MEMBRANE FUSION', i.e. Result(e16,s4), we need detailed knowledge about the exocytosis pathway and how membrane fusion could be potentially related to it.

The resolution of bridging references like the previous one are clearly out of reach for now because they involve too detailed domain knowledge about biochemical pathways and their mechanisms as well as their relations to other ones. In fact the author sees no way of significantly improving the recall of the approach without a detailed model of the general mechanism of signaling or metabolic pathways as well as detailed knowledge about specific pathways.

For a more detailed analysis of the precision yielded by the approach, two further biologists have been asked to go through the discourse relations which have been classified as incorrect when evaluating the results on the test corpus measured against the 'ALL' standard and to check if any of them actually represents a valid interpretation. In this sense, an answer given by the system has been *a posteriori* regarded as correct if at least one of these subjects considered it as corresponding to a valid interpretation. As a result, 8 further discourse relations have been identified as actually valid thus yielding an improved precision of P=90.98%. The remaining 9% erroneous answers can clearly be traced back to generalizations undertaken within the ontology development step. These generalizations have been intended to increase recall without the need of writing rules for each special case, and the price to pay for this economy of rules seems to be a reduction in precision.

8.6 Exploiting Lexical Clues in the Resolution Process

A further interesting observation is that in many cases there are lexical clues which already indicate the conceptual relation between two propositions or events. This is in particular the case for conjunctions such as *by*, *thus*, *because*, *also*, just to name a few. This observation has led to the idea that discourse relations could also be lexically inferred. For this purpose, the semantic representation of the text has been enriched with a predicate specifying the lexical element by which events are connected. Take the following example:

- (80) ALPHA-CONOTOXINS <id="s42" ACT> ON POSTSYNAPTIC
 MEMBRANES, THEY <id="b53" BIND> TO <id="d85" THE
 NICOTINIC ACETYLCHOLINE RECEPTORS (NACHR)> AND
 THUS <id="e126" INHIBIT> THEM.

Conjunction	Discourse Relation
after	Narration
also	Narration
because	Explanation/Elaboration
but	Contrast
by	Explanation/Elaboration
then	Narration
thereby	Result
therefore	Result
through	Explanation/Elaboration
thus	Result
via	Explanation/Elaboration

Table 12: Conjunctions and the corresponding discourse relations inferred from them

The connection between the events *b53* and *e126* would then in the sense above be represented by the predicate: *connection(b53, e126, 'thus')*. On the basis of such a representation we could now define rules for example stating that if two events are lexically connected via the conjunction *thus*, then normally *Result* is the relation between them, i.e.

$$\forall e_1, e_2 \text{ connection}(e_1, e_2, 'thus') \rightarrow \text{Result}(e_1, e_2) \quad (81)$$

In this sense, a lexicon containing the conjunctions appearing in the test corpus as well as the corresponding discourse relation which can be 'lexically' inferred from them has been built (see table 12). Then the test corpus has been annotated with the *connection* predicate using the conjunctions specified in this lexicon. Furthermore, a simple *Perl*-program has been developed which given a specific *connection*-predicate specifying the conjunction linking two events together, infers the corresponding discourse relation from table 12. This 'lexically driven approach', as it will be referred to, yielded a very high precision on the test corpus (100%) but very low recall values measured on the three standards (16.23% - 18.18%). These results suggest that most of the discourse relations in the corpus can not be inferred by lexical means and shows that a knowledge-based approach is in fact necessary. Nevertheless the results also suggest that the ontology and lexically driven approaches could be combined somehow to increase the performance of the whole bridging resolution process. Figure 9 outlines graphically how such a combination might work. It furthermore shows that the decisive question is how to combine the set A of 'ontologically inferred' and the set B of 'lexically inferred' discourse relations. In fact, taking into account the low recall of the lexically driven approach, it seems obvious that the set A will basically determine the overall recall of the system while B will be

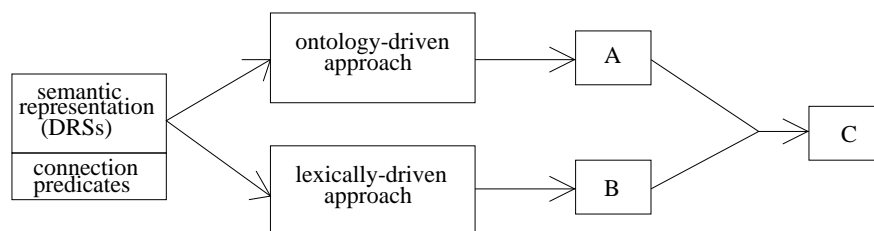


Figure 9: Combination of the ontology-driven and the lexically-driven approach

	AUTHOR			2/3			ALL		
	R	P	F	R	P	F	R	P	F
1	9.2%	100%	16.84%	7.79%	100%	14.45%	12.12%	100%	21.62%
2	61.41%	93.19%	74.03%	48.70%	93.19%	63.97%	63.63%	93.19%	75.62%
3	53.84%	91.67%	67.84%	38.31%	91.67%	54.04%	54.54%	91.67%	68.39%

Table 13: Results of the combination of the ontology-driven and the lexically driven approach

responsible for increasing the overall precision by eliminating incorrect readings from A. So three ways of combining both sets thus producing a final set C seem possible:

1. $A \cap B$
2. $A \cup B - inconsistent(A, B)$
3. $A - inconsistent(A, B)$

where $inconsistent(A, B)$ is the set of elements of A and B which given a certain referring expression differ in the corresponding discourse relation between this expression and some antecedent. Table 13 shows the results of these three combination strategies evaluated against the three standards. It clearly shows that combination strategy 2 significantly increases not only the precision but also the recall of the whole approach. The conclusion is that the lexically driven approach outlined in this section can in fact complement the ontology-driven approach and significantly improve the overall performance of the system.

8.7 The Turing-test

In ([68]), Turing proposed a test to evaluate the intelligence of a machine or a system. The scenario of this test is as follows: one person sitting in front of a

Set	Kappa coefficient
(a_1, a_2, a_3)	0.31
(s, a_2, a_3)	0.32
(a_1, s, a_3)	0.45
(a_1, a_2, s)	0.46

Table 14: Agreement of the sets defined in 21

computer terminal and communicates with another person at another computer terminal as well as with a machine. The latter two are at a different place so that the former person can not see any of them. The test now consists in identifying which communication partner is the machine and which is the human subject by conversating with both of them. Turing's idea when defining this test was to establish a criterion in order to regard a machine or a system as intelligent provided it can not be distinguished from the human subject. If we abstract from the dialogue task envisioned by Turing, this criterion boils down to the following one:

Definition 20 (Success of a system)

A system s will be regarded as intelligent or successful on a task x if the output o of the system on x ($o_s(x)$) can not be distinguished from the one of an arbitrary human subject.

When applying these ideas to the task of resolving bridging references, a test to measure the success of a system on this task can be defined as follows:

Definition 21 (Success on the bridging reference resolution task)

Given a set of annotators a_1, \dots, a_n identifying bridging references in a certain corpus and a system s performing the same task automatically as well as the agreement between 'subjects' in the sets $\{a_1, \dots, a_n\}$, $\{s, a_2, \dots, a_n\}$, $\{a_1, \dots, a_{n-1}, s\}$ and $\{a_1, \dots, a_{j-1}, s, a_{j+1}, \dots, a_n\}$ with $j=2..n-1$, the system will be considered as succesful if on the basis of the agreement of these sets it is not possible to identify the set in which the system s has not been considered, i.e. if the output of the system does not significantly differ from the one of the other human subjects.

In the special case presented within this diploma thesis, the number of annotators is three : a_1 , a_2 and a_3 . The agreement as measured by the Kappa statistic of the sets defined in 21 is indicated in table 14. The measures show that there was a particular good agreement between the annotator a_1 and the system s , but also that it is clearly impossible to identify the set in which the system has not been considered only on the basis of the agreement measures. So following

	Recall	Precision	F-measure
AUTHOR	61.41%	93.19%	74.03%
2/3	48.70%	93.19%	63.97%
ALL	63.63%	93.19%	75.62%

Table 15: Final Results: ontology-driven approach and lexically-driven approach combined

definition 21 the system can be definitely regarded as succesful on the bridging reference resolution task.

8.8 Conclusions

It has turned out that the agreement of different annotators on the task of identifying relations between events as antecedents and other events, states or entities as referring expressions is not very high. This suggests that the task itself is not a trivial one and that it is highly subjective. Consequently the performance of a machine-based approach to bridging reference resolution involving events as antecedents has its limits and can not be expected to show a very high performance.

The performance of the approach described in section 4 has been evaluated against a 'gold standard' developed on the basis of the annotations of three different subjects. The recall of the approach is $R=45.38\%$ when considering the relations identified by at least two of the three annotators and $R=54.54\%$ when considering only the ones on which the three annotators agree. The precision is $P=90.98\%$ and has been measured against the union of the relations found by all of the annotators. Otherwise the approach would be penalized for finding relations which are actually correct. In determining the precision it has also been considered that for some bridging references the antecedents are ambiguous.

Furthermore, the results presented in section 8.6 show that the successful combination of the ontology-driven approach with a mechanism to lexically infer discourse relations can improve the overall recall and precision of the system. Table 15 presents the final results corresponding to the combination of both approaches as outlined in section 8.6. However, these results are not directly comparable to the ones of the systems discussed in section 3.1.5 and 3.2 due to several reasons. First, the domain of application is different from the one of all the other systems. Second, most systems concentrate on the resolution of coreferences between objects or events but none of them attempts to compute discourse relations between events, so that the task at hand seems inherently harder. Third, the approach presented here has been evaluated given a semantic representation of the text, while the other systems have been evaluated either given a syntactic representation or even raw text. Nevertheless, a comparison

Poesio et al.			Muñoz et al.			Hahn et al.		
R	P	F	R	P	F	R	P	F
57%	70%	62.83%	75.3%	79.3%	77.25%	55.05%	95.2%	69.76%
Circus (MUC6)			Circus (MUC7)			Proteus (MUC6)		
R	P	F	R	P	F	R	P	F
50.71%	71.93%	59.48%	56.1%	68.80%	61.8%	53%	62%	57.15%

Table 16: Results of the systems discussed in section 3.1.5 and 3.2

between the results of the approach presented here and the ones in table 16 shows that it fits quite well in the picture of the results of other systems dealing with a similar task in the field of discourse analysis.

9 Conclusion and Outlook

This master's thesis has on the one hand presented an ontology-driven approach to bridging reference resolution within information extraction systems. The approach makes use of a semantic representation of the text to make contextual information explicit as well as of a model of the domain in form of an ontology to infer conceptual relations between events. In particular it has been shown that by considering not only definite descriptions but also verbs representing events or states as referring expressions thus being subject to the bridging reference resolution process, discourse relations between events as antecedents and these expressions can be computed. This assumption has not been linguistically motivated nor discussed. In fact it would be interesting to analyze from a theoretical as well as an empirical point of view in how far this assumption can be regarded as linguistically sound.

The strategy presented to find a suitable antecedent for a referring expression has been basically to 'go back' in the discourse processed so far and infer a suitable antecedent for this expression as well as the potential relation between them. However, it is certainly also possible to model the presuppositions introduced by the referring expression and then go back in the discourse and find a suitable antecedent for the presupposed information. In fact, though only the first possibility has been explored within the work presented here, the author's conclusion is that the combination of both methods is the key to a successful resolution of bridging references. For this purpose, it would be necessary to explore this second possibility as well as the principles under which both strategies could be combined.

On the other hand, a lexically-driven approach has been presented which on the basis of a given conjunction linking together two events, allows to infer a specific discourse relation between them. It has also been discussed how the ontology-driven approach and the latter one can be combined to yield better overall results in terms of recall and precision. The author would like to point out that such a lexically-driven method to infer rhetorical relations 'bottom-up' (alla DICE) can be used for lexical and in consequence also for semantic disambiguation as discussed by Asher et. Lascarides ([1]).

Furthermore it has turned out that annotating discourse relations between events and states is not a trivial task. The low agreement between annotators shows that it is actually a very subjective one. An interesting observation is that in some cases the antecedent for a specific event is ambiguous. However, as the annotators were not allowed to underspecify the antecedent, they were forced to disambiguate and in fact chose different antecedents for these ambiguous cases. This suggests that in these cases the choice of the antecedent does not affect the (semantic) interpretation of the text. The author has hypothesized that in the examples in question this could be due to the transitivity of causality and not due to an underlying mereological structure as presented in [50]. Regarding this question it would be very interesting to conduct experiments with minimal pairs to find out if the transitivity of causality is actually the condition which allows a certain antecedent to be ambiguous without affecting the overall

interpretation. An even more interesting question is whether subjects are really 'sloppy' in the sense that they do not disambiguate the antecedent if they are not forced to or if they actually choose a concrete antecedent. It would also be interesting to know if they are aware of the fact that their choice or sloppiness does not affect the interpretation.

This thesis has in addition shown that it is possible to classify the verbs within the domain in question into a few (non-disjoint) classes and identify the common and characteristic pre- and postconditions of the events represented by the verbs of each class. Furthermore the principles have been explored under which, on the basis of a taxonomy of these pre- and postconditions and the above mentioned classification, a taxonomy of biochemical events can be developed. Such a taxonomy can then be used to model the necessary conceptual relations between events at the highest possible level of abstraction.

The thesis has also presented a formal definition of an ontology as well as an inference mechanism to reason within it. It is also important to mention that if an ontology is regarded as a logical theory ([19]), the inference mechanism presented is correct but not complete with regard to this theory. In fact it would be interesting to explore more refined inference mechanisms and integrate them into the bridging reference resolution algorithm. For this purpose the general role of world knowledge within anaphora resolution has been investigated and an abstract and generalized version of the bridging reference resolution of Bos et al. has been presented allowing any form of knowledge representation formalism and a corresponding inference mechanism to be integrated. In particular, the author finds the idea of Blackburn et al. to run several theorem provers in parallel an interesting and promising one ([5]).

It should be stressed that the approach developed is neither inherently restricted to DRT as discourse representation nor as knowledge representation language. In fact the only restriction on an alternative discourse representation formalism is that a notion of homomorphism can be defined on the corresponding structures. On the other hand the only restriction concerning the use of another knowledge representation formalism is that an inference mechanism can be defined on it and that it can be translated to the discourse representation formalism used.

The approach developed in this thesis has been successfully applied to a training and test corpus of SWISS-PROT function descriptions. However, the approach can definitely be further developed and refined. For example, it would be interesting not only to consider discourse relations between events but also mereological or spatial relations between objects mentioned in the function descriptions thus also making these relations explicit within the bridging reference resolution process. For this purpose it could turn out to be important to introduce a semantic representation not only for the DE line ([3]) but also for other lines containing mereological information about the protein in question (CC SUBUNIT), the gene it is synthesized from (GN line), the organism in which it is found (OS and OC lines), the pathways it is involved in (CC PATHWAY), its subcellular location (CC SUBCELLULAR LOCATION) etc. as context for the interpretation of the function description of the protein.

Moreover, the work presented here has also shown that a more detailed model of pathways in general as well as of specific pathways and their relations is needed to capture many of the conceptual relations between events. In this sense it would be interesting to concentrate on and model in detail one specific pathway and apply the bridging reference resolution to texts dealing with this restricted domain in order to verify if through the computation of discourse relations new relations with regard to the model can be found.

The approach as well as the ontology have been developed in a general way so that in principle they are also applicable to other texts of a biochemical nature. However, it should be mentioned that the approach has only been tested on SWISS-PROT function descriptions which are relatively short so that, when applying it to larger texts, the system's performance could be significantly slower. Certainly the approach is general and flexible enough to be applied to other domains. However, for this purpose a suitable ontology of events and their conceptual relations would have to be developed for the domain in question. So is the system developed in fact dumb as it can only perform on one domain? Well, probably it is like with people too:

Everybody's ignorant, only on different subjects. (Will Rogers)

A Ambiguous Antecedents

This appendix contains examples taken out of the test corpus for which the annotators agree in the referring expression as well as in the (rhetorical) relation connecting it to a certain antecedent but differ in the choice of this antecedent which thus has to be regarded as ambiguous. In the following examples this ambiguity in the choice of the antecedent is expressed by indicating the possible antecedents in square brackets. If for example an event e can be interpreted as a result of either an event e_1 or an event e_2 , this will be represented as $\text{Result}([e_1, e_2], e)$:

<Names> Coatomer gamma-2 subunit (Gamma-2 coat protein) (Gamma-2 COP). </Names>
<FUNCTION> <id="d67" THE COATOMER> IS A CYTOSOLIC PROTEIN COMPLEX THAT <id="b44" BINDS> TO DILYSINE MOTIFS AND REVERSIBLY <id="e99" ASSOCIATES> WITH GOLGI NON-CLATHRIN-COATED VESICLES, WHICH FURTHER <id="e100" MEDIATE> BIOSYNTHETIC PROTEIN <id="e101" TRANSPORT> FROM <id="d68" THE ER>, VIA <id="d69" THE GOLGI> UP TO <id="d70" THE TRANS GOLGI NETWORK>. COATOMER COMPLEX IS <id="s35" REQUIRED> FOR <id="e102" BUDDING> FROM GOLGI MEMBRANES, AND IS ESSENTIAL FOR <id="d71" THE RETROGRADE GOLGI-TO-ER <id="e103" TRANSPORT> OF DILYSINE-TAGGED PROTEINS>. IN MAMMALS, <id="d72" THE COATOMER> CAN ONLY BE <id="e104" RECRUITED> BY MEMBRANES <id="s36" ASSOCIATED> TO ADP-RIBOSYLATION FACTORS (ARFS), WHICH ARE SMALL GTP-BINDING PROTEINS; <id="d73" THE COMPLEX> ALSO <id="e105" INFLUENCES> <id="d74" THE GOLGI STRUCTURAL INTEGRITY>, AS WELL AS <id="e106" THE PROCESSING, ACTIVITY, AND ENDOCYTIC RECYCLING OF LDL RECEPTORS> (BY SIMILARITY). </FUNCTION>
 $\text{Result}([b44, e99], e100)$

<Names> DNA-binding protein HU. </Names>
<FUNCTION> <id="d92" THIS PROTEIN> <id="s45" BELONGS> TO <id="d93" THE HISTONE LIKE FAMILY OF PROKARYOTIC <id="b270" DNA-BINDING> PROTEINS WHICH ARE CAPABLE OF <id="e132" WRAPPING> DNA TO <id="e133" STABILIZE> IT, AND <id="e134" PREVENT> ITS DENATURATION UNDER EXTREME ENVIRONMENTAL CONDITIONS>. ESSENTIAL FOR HETEROCYST DIFFERENTIATION. </FUNCTION>
 $\text{Result}([b270, e132], e133)$

<Names> Eukaryotic translation initiation factor 2 alpha subunit (eIF-2- alpha). </Names>
<FUNCTION> EIF-2 <id="s91" FUNCTIONS> IN <id="d168" THE EARLY STEPS OF PROTEIN SYNTHESIS> BY <id="e231" FORMING> A TERNARY COMPLEX WITH GTP AND INITIATOR TRNA. <id="d169" THIS COMPLEX>

<id="b106" BINDS> TO A 40S RIBOSOMAL SUBUNIT, <id="e232" FOLLOWED> BY MRNA <id="b107" BINDING> TO <id="e233" FORM> A 43S PREINITIATION COMPLEX. <id="e618" JUNCTION> OF <id="d170" THE 60S RIBOSOMAL SUBUNIT> TO <id="e234" FORM> <id="d171" THE 80S INITIATION COMPLEX> IS <id="e235" PRECEDED> BY HYDROLYSIS OF <id="d172" THE GTP <id="s92" BOUND> TO EIF-2> AND <id="e236" RELEASE> OF AN EIF-2-GDP BINARY COMPLEX. IN ORDER FOR EIF-2 TO <id="e237" RECYCLE> AND <id="e238" CATALYZE> ANOTHER ROUND OF <id="e239" INITIATION>, <id="d173" THE GDP <id="s93" BOUND> TO EIF-2> MUST <id="e240" EXCHANGE> WITH GTP BY WAY OF A <id="e241" REACTION> <id="e242" CATALYZED> BY EIF-2B. </FUNCTION>
Result([b106,b107],e233)

<Names> Arginine deiminase (EC 3.5.3.6) (ADI) (Arginine dihydrolase) (AD). </Names>
<FUNCTION> MEMBER OF <id="d16" THE TWO-COMPONENT REGULATORY SYSTEM ARCB/ARCA>. <id="e27" REPRESSES> A WIDE VARIETY OF AEROBIC ENZYMES UNDER ANAEROBIC CONDITIONS. <id="s7" CONTROLS> <id="s8" THE RESISTANCE OF E.COLI TO DYES>; <id="s9" REQUIRED> FOR <id="e28" EXPRESSION> OF <id="d17" THE ALKALINE PHOSPHATASE AND SEX FACTOR F GENES>; IT ALSO MAY BE <id="s10" INVOLVED> IN <id="e29" THE OSMOREGULATION OF ENVELOPE PROTEINS>. WHEN <id="e30" ACTIVATED> BY ARCB, IT NEGATIVELY <id="e31" REGULATES> <id="e32" THE EXPRESSION OF GENES OF AEROBIC FUNCTION>. <id="e33" ACTIVATES> <id="e34" THE TRANSCRIPTION OF <id="d18" THE PLFB OPERON>> BY <id="b14" BINDING> TO ITS PROMOTER. </FUNCTION>
Result([e27,e30],e31)

<Names> Regulator of G-protein signaling 19 (RGS19). </Names>
<FUNCTION> <id="e423" INHIBITS> SIGNAL TRANSDUCTION BY <id="e424" INCREASING> <id="e425" THE GTPASE ACTIVITY OF G PROTEIN ALPHA SUBUNITS> THEREBY <id="e426" DRIVING> THEM INTO THEIR INACTIVE GDP-BOUND FORM. <id="b189" BINDS> TO G-ALPHA SUBFAMILY 1 MEMBERS, WITH <id="d329" THE ORDER G(I)A3 > G(I)A1 > G(O)A >> G(Z)A/G(I)A2> <id="e427" ACTIVITY> ON G(Z)-ALPHA IS <id="e428" INHIBITED> BY <id="e429" PHOSPHORYLATION> AND <id="e430" PALMITOYLATION> OF <id="d330" THE G-PROTEIN> (BY SIMILARITY). </FUNCTION>
Result([e423,e424],e426)

<Names> Apoptosis regulator BAX, membrane isoform alpha. </Names>
<FUNCTION> <id="e41" ACCELERATES> PROGRAMED CELL DEATH BY <id="b18" BINDING> TO, AND <id="e42" ANTAGONIZING> <id="d22" THE APOPTOSIS REPRESSOR BCL-2> OR ITS ADENOVIRUS HOMOLOG

E1B 19K PROTEIN. <id="e43" INDUCES> <id="e44" THE RELEASE OF CYTOCHROME C>, <id="e45" ACTIVATION> OF CASPASE-3, AND THEREBY APOPTOSIS. </FUNCTION>

Result([e43,b18],e45)

<Names> Neurotoxin II (BT-II) (Fragment). </Names>

<FUNCTION> <id="b208" BINDS> TO SODIUM CHANNELS AND <id="e457" INHIBITS> <id="e458" THE INACTIVATION OF <id="d369" THE <id="s174" ACTIVATED> CHANNELS>>, THEREBY <id="e459" BLOCKING> NEURONAL TRANSMISSION. <id="d370" THIS TOXIN> IS ACTIVE AGAINST MAMMALS. LD(50) IS 2.25 MG/KG IN MICE BY SUBCUTANEOUS INJECTION. </FUNCTION>

Result([b208,e457],e459)

<Names> Ras-related protein SEC4. </Names>

<FUNCTION> <id="s175" INVOLVED> IN EXOCYTOSIS. MAYBE BY <id="e460" REGULATING> <id="d371" THE <id="b209" BINDING> AND <id="e461" FUSION> OF SECRETORY VESICLES WITH <id="d372" THE CELL SURFACE>>. <id="d373" THE GTP-BOUND FORM OF SEC4> MAY <id="e462" INTERACT> WITH AN EFFECTOR, THEREBY <id="e463" STIMULATING> ITS <id="e464" ACTIVITY> AND <id="e465" LEADING> TO EXOCYTOTIC <id="e466" FUSION>. SEC4 MAY BE AN UPSTREAM ACTIVATOR OF <id="d374" THE 19.5S SEC8/SEC15 PARTICLE>. SEC4 PROBABLY <id="e467" INTERACTS> DIRECTLY WITH SEC8; IT COULD <id="s176" SERVE> AS <id="d375" THE ATTACHMENT SITE FOR <id="d376" THE SEC8/SEC15 PARTICLE>> (BY SIMILARITY). </FUNCTION>

Result([e463,e462],e465)

<Names> Insect toxin 1 (BsIT1) (Bs-dprIT1). </Names>

<FUNCTION> <id="b214" BINDS> TO SODIUM CHANNELS AND <id="e480" INHIBIT> <id="e481" THE INACTIVATION OF <id="d383" THE <id="s179" ACTIVATED> CHANNELS>>, THEREBY <id="e482" BLOCKING> NEURONAL TRANSMISSION. <id="d384" THIS DEPRESSANT INSECT TOXIN> <id="e483" DEPOLARIZES> <id="d385" THE COCKROACH AXON>, IRREVERSIBLY <id="e484" BLOCKS> <id="d386" THE ACTION POTENTIAL>, AND <id="e485" SLOWS> DOWN AND VERY PROGRESSIVELY <id="e486" BLOCKS> <id="d387" THE TRANSMEMBRANE TRANSIENT SODIUM CURRENT>. </FUNCTION>

Result([b214,e480],e482)

Result([e485,e483],e486)

<Names> Transposable element TC3 transposase. </Names>

<FUNCTION> <id="b237" BINDS> SPECIFICALLY TO <id="d438" THE TERMINAL NUCLEOTIDES OF <id="d439" THE TC3 INVERTED REPEAT>>. ITS <id="e527" EXPRESSION> <id="e528" RESULTS> IN FREQUENT <id="e529" EXCISION> AND <id="e530" TRANSPOSITION> OF ENDOGE-

NOUS TC3 ELEMENTS. TC3 TRANSPOSASE <id="s198" ACTS> BY MAKING DOUBLE STRAND BREAKS AT <id="d440" THE ENDS OF TC3 ELEMENT>. <id="d441" THE <id="e531" EXCISED> ELEMENT> WOULD THEN BE <id="e532" INSERTED> INTO A TARGET SEQUENCE. </FUNCTION>
Result([e528,e527],e529)
Result([e528,e527],e530)

<Names> Cyclin-dependent kinase inhibitor 1 (Melanoma differentiation associated protein 6) (MDA-6) (P21) (CDK-interacting protein 1). </Names>
<FUNCTION> MAY BE <id="d56" THE IMPORTANT INTERMEDIATE> BY WHICH P53 <id="e86" MEDIATES> ITS ROLE AS AN INHIBITOR OF CELLULAR <id="e615" PROLIFERATION> IN RESPONSE TO DNA DAMAGE. MAY <id="b37" BIND> TO AND <id="e87" INHIBIT> CYCLIN-DEPENDENT KINASE <id="e88" ACTIVITY>, <id="e89" PREVENTING> <id="e90" PHOSPHORYLATION> OF CRITICAL CYCLIN-DEPENDENT KINASE SUBSTRATES AND <id="e91" BLOCKING> CELL CYCLE PROGRESSION. </FUNCTION>
Result([b37,e87],e89)

<Names> Gamma-aminobutyric-acid receptor alpha-1 subunit precursor (GABA(A) receptor). </Names>
<FUNCTION> GABA, <id="d129" THE MAJOR INHIBITORY NEUROTRANSMITTER IN <id="d130" THE VERTEBRATE BRAIN>>, <id="e181" MEDIATES> NEURONAL <id="e182" INHIBITION> BY <id="b84" BINDING> TO <id="d131" THE GABA/BENZODIAZEPINE RECEPTOR> AND <id="e183" OPENING> AN INTEGRAL CHLORIDE CHANNEL. </FUNCTION>
Explanation/Elaboration([e182,e181],b84)

<Names> Gamma-aminobutyric-acid receptor rho-1 subunit precursor (GABA(A) receptor). </Names>
<FUNCTION> GABA, <id="d132" THE MAJOR INHIBITORY NEUROTRANSMITTER IN <id="d133" THE VERTEBRATE BRAIN>>, <id="e184" MEDIATES> NEURONAL <id="e185" INHIBITION> BY <id="b85" BINDING> TO <id="d134" THE GABA/BENZODIAZEPINE RECEPTOR> AND <id="e186" OPENING> AN INTEGRAL CHLORIDE CHANNEL. RHO-1 GABA RECEPTOR COULD <id="s66" PLAY> A ROLE IN RETINAL NEUROTRANSMISSION. </FUNCTION>
Explanation/Elaboration([e184,e185],b85)

<Names> Erythroid protein 4.1 (Band 4.1) (Fragment). </Names>
<FUNCTION> PROTEIN 4.1 IS A MAJOR STRUCTURAL ELEMENT OF <id="d1" THE ERYTHROCYTE MEMBRANE SKELETON>. IT <id="s1" PLAYS> A KEY ROLE IN <id="e1" REGULATING> MEMBRANE PHYSICAL PROPERTIES OF MECHANICAL STABILITY AND DEFORMABILITY BY <id="e2" STABILIZING> <id="e3" SPECTRIN-ACTIN INTERACTION>. <id="b1" BINDS> WITH A HIGH AFFINITY TO GLYCOPHORIN AND

WITH LOWER AFFINITY TO BAND III PROTEIN. </FUNCTION>
Explanation/Elaboration([s1,e1],e2)

<Names> Methyl-CpG-binding protein 2 (MeCP-2 protein) (MeCP2). </Names>
<FUNCTION> CHROMOSOMAL PROTEIN THAT <id="b133" BINDS>
TO <id="s109" METHYLATED> DNA. IT CAN <id="b134" BIND> SPECIF-
ICALLY TO A SINGLE METHYL-CPG PAIR. IT IS NOT <id="e293" INFLUENCED>
BY SEQUENCES <id="s110" FLANKING> <id="d231" THE METHYL-
CPGS>. <id="e294" MEDIATES> TRANSCRIPTIONAL <id="e295" REPRESSION>
THROUGH <id="e296" INTERACTION> WITH HISTONE DEACETYLASE
AND <id="d232" THE COREPRESSOR SIN3A>. </FUNCTION>
Explanation/Elaboration([e295,e294],e296)

<Names> Sialidase precursor (EC 3.2.1.18) (Neuraminidase) (NANASE). </Names>
<FUNCTION> <id="e315" CLEAVES> <id="d250" THE TERMINAL SIALIC
ACID (N-ACETYL NEURAMINIC ACID) FROM CARBOHYDRATE CHAINS
IN GLYCOPROTEINS> <id="e316" PROVIDING> FREE SIALIC ACID
WHICH CAN BE <id="e317" USED> AS CARBON AND ENERGY SOURCES.
SIALIDASES HAVE BEEN SUGGESTED TO BE PATHOGENIC FACTORS
IN MICROBIAL INFECTIONS. NANH <id="e318" FACILITATES> CHOLERA
TOXIN <id="b147" BINDING> TO HOST INTESTINAL EPITHELIAL CELLS
BY <id="e319" CONVERTING> CELL SURFACE POLYSIALOGLANGLI-
OSIDES TO GM1 MONOGLANGLIOSIDES. </FUNCTION>
Explanation/Elaboration([b147,e318],e319)

<Names> Twist related protein (H-twist). </Names>
<FUNCTION> PROBABLE TRANSCRIPTION FACTOR, WHICH SEEMS
TO BE <id="s206" INVOLVED> IN <id="d466" THE NEGATIVE REGU-
LATION OF CELLULAR DETERMINATION> AND IN <id="e560" THE
DIFFERENTIATION OF SEVERAL LINEAGES INCLUDING MYOGENE-
SIS, OSTEOGENESIS, AND NEUROGENESIS>. <id="e561" INHIBITS>
MYOGENESIS BY <id="e562" SEQUESTERING> E PROTEINS, <id="e563"
INHIBITING> TRANS-ACTIVATION BY MEF2, AND <id="e564" INHIBITING>
<id="b255" DNA-BINDING> BY MYOD THROUGH PHYSICAL <id="e565"
INTERACTION>. <id="e566" THIS INTERACTION> PROBABLY <id="s207"
INVOLVES> <id="d467" THE BASIC DOMAINS OF BOTH PROTEINS>
(BY SIMILARITY). </FUNCTION>
Explanation/Elaboration([s206,e561],e564)

<Names> Regulatory protein E2. </Names>
<FUNCTION> E2 <id="e585" REGULATES> VIRAL <id="e586" TRANSCRIPTION>
AND <id="e587" DNA REPLICATION>. IT <id="b261" BINDS> TO <id="d473"
THE E2RE RESPONSE ELEMENT (5'-ACCNNNNNNGGT-3') PRESENT
IN MULTIPLE COPIES IN <id="d474" THE REGULATORY REGION>>.
IT CAN EITHER <id="e588" ACTIVATE> OR <id="e589" REPRESS>
<id="e590" TRANSCRIPTION> DEPENDING OF E2RE'S POSITION WITH

REGARDS TO PROXIMAL PROMOTER ELEMENTS. <id="e591" REPRESSION>
<id="e592" OCCURS> BY STERICALLY <id="e593" HINDERING> <id="e594"
THE ASSEMBLY OF <id="d475" THE TRANSCRIPTION INITIATION COMPLEX>>.
<id="s210" THE E1-E2 COMPLEX> <id="b262" BINDS> TO <id="d476"
THE ORIGIN OF DNA REPLICATION>. </FUNCTION>
Explanation/Elaboration([e592,e591],e593)

<Names> Chromatin assembly factor 1 P55 subunit (CAF-1 P55 subunit)
(DCAF-1) (Nucleosome remodeling factor 55 kDa subunit) (NURF-55). </Names>
<FUNCTION> COMPLEX THAT <id="e62" ASSEMBLES> HISTONE OC-
TAMERS ONTO REPLICATING DNA IN VITRO. CAF-1 <id="e63" PERFORMS>
<id="d39" THE FIRST STEP> OF <id="d40" THE NUCLEOSOME AS-
SEMBLY PROCESS>, <id="e64" BRINGING> NEWLY <id="s15" SYN-
THESIZED HISTONES H3 AND H4 TO REPLICATING DNA; HISTONES
H2A/H2B CAN <id="b27" BIND> TO <id="d41" THIS CHROMATIN PRECURSOR>
SUBSEQUENT TO DNA <id="e65" REPLICATION> TO <id="e66" COMPLETE>
<id="d42" THE HISTONE OCTAMER>. P150 AND P60 <id="e67" FORM>
COMPLEXES WITH NEWLY <id="s16" SYNTHESIZED> HISTONES H3
AND <id="s17" ACETYLATED> H4 IN CELL EXTRACTS (BY SIMILAR-
ITY). P55 <id="e68" ASSOCIATES> WITH ACETYLTRANSFERASES AND
DEACETYLASES IN CELL EXTRACTS AND IS ALSO A COMPONENT
OF <id="d43" THE NUCLEOSOME REMODELING FACTOR COMPLEX
(NURF)>, A PROTEIN COMPLEX <id="e69" CONSISTING> OF FOUR
POLYPEPTIDES THAT <id="e70" FACILITATES> <id="d44" THE PER-
TURBATION OF CHROMATIN STRUCTURE> IN VITRO IN AN ATP-
DEPENDENT MANNER. <id="d45" THIS SUBUNIT> MAY <id="e71" ASSIST>
<id="e72" TARGETING> OF PROTEIN COMPLEXES TO CHROMATIN.
</FUNCTION>
Explanation/Elaboration([e63,e62],e64)

B SWI Prolog Source Code

B.1 anaphora_resolution.pl

```
resolve(X) :- projectDRS(X,[],Res), inverse(Res,Inv), print(Inv), nl ,fail.

projectDRS([merge(A1,A2)],K,Res) :- projectDRS([A1],K,K2),
                                   projectDRS([A2],K2,Res).

projectDRS([alfa(drs(D,C))],K,Res) :- projectConds(C,K,K2,C2,R1),
                                   ar(drs(D,C2),K2,K3,R2),
                                   append(C2,R1,C4),
                                   add(C4,R2,C5),
                                   append([drs(D,C5)],K3,Res).

projectDRS([drs(D,C)],K,Res) :- projectConds(C,K,K2,R,B),
                                append(C,R,C2),
                                add(C2,B,C3),
                                append([drs(D,C3)],K2,Res).

projectConds([],K,K,[],[]).

projectConds([alfa(drs(D,C))|R1],K,K4,Res,B5) :- projectConds(C,K,K2,R,B1),
                                                ar(drs(D,R),K2,K3,B2),
                                                append(B1,B2,B3),
                                                projectConds(R1,K3,K4,R3,B4),
                                                append(R,R3,Res),
                                                append(B3,B4,B5), !.

projectConds([or(D1,D2)|R1],K,K4,R3,B5) :- projectConds([D1],K,K2,D3,B1),
                                                projectConds([D2],K2,K3,D4,B2),
                                                append(B1,B2,B3),
                                                projectConds(R1,K3,K4,R2,B4),
                                                append([or(D3,D4)],R2,R3),
                                                append(B3,B4,B5), !.

projectConds([drs(D,C)|R1],K,K3,[drs(D,C2)|R2],B3) :- projectConds(C,K,K2,C2,B1),
                                                        projectConds(R1,K2,K3,R2,B2),
                                                        append(B1,B2,B3), !.

projectConds([X|R1],K,K2,[X|R2],B) :- projectConds(R1,K,K2,R2,B).

ar(A,K,K2,R) :- linking(A,K,K2,R) ,!.
ar(A,K,K2,R) :- bridging(A,K,K2,R) ,!.
ar(A,[D|K1],[D|K2],B) :- ar(A,K1,K2,B), !.
ar(A,K,K2,R) :- accomodate(A,K,K2,R).
```

```

m_suitable([K1|C],C2,B,M) :- member(K2,C2,C3),
                             K1 =.. [=,A1,A2],
                             K2 =.. [=,A3,A4],
                             unify(A2,A4),
                             member(A1=A3,B,_),
                             m_suitable(C,C3,B,M).

```

```

m_suitable([K1|C],C2,B,M) :- member(K2,C2,C3),
                             K1 =.. [=,A1,A2],
                             K2 =.. [=,A3,A4],
                             unify(A2,A4),
                             \+ member(_=A3,B,_),
                             append(A1=A3,B,L),
                             m_suitable(C,C3,L,M).

```

```

m_suitable([K1|C],C2,B,M) :- member(K2,C2,C3),
                             K1 =.. [P1|R1],
                             K2 =.. [P2|R2],
                             P1 = P2,
                             \+ unify(P1,=),
                             \+ unify(P1,drs),
                             equal(R1,R2,B,L),
                             m_suitable(C,C3,L,M).

```

```

m_suitable([],_,B,B).

```

```

member(X,[Y|R],[Y|Res]) :- member(X,R,Res).
member(X,[X|R],R).

```

```

equal([X|R1],[Y|R2],B,R) :- member(X=Y,B,_), equal(R1,R2,B,R).
equal([X|R1],[Y|R2],B,[X=Y|R]) :- \+ member(X=_,B,_), equal(R1,R2,B,R).
equal([],[],B,B).

```

```

unify([],[]) :- !.
unify(X,X) :- !.
unify(_,'?') :- !.
unify('?',_) :- !.
unify([X1|R1],R) :- member(X1,R,R2), unify(R1,R2).

```

```

add([X|R1],L,Res) :- member(X,L,_), add(R1,L,Res), !.
add([X=X|R1],L,Res) :- add(R1,L,Res), !.
add([X|R1],L,[X|Res]) :- add(R1,L,Res).
add([],L,L) :- !.

```

B.2 bridging.pl

```
linking(drs(.,C1),[drs(D2,C2)|K],[drs(D2,C2)|K],M) :- m_suitable(C1,C2,[],M), !.

bridging(drs(.,C1),[drs(D2,C2)|K],[drs(D2,C2)|K],M) :- is_a*(drs(D2,C2),drs(.,C3)),
    m_suitable(C1,C3,[],M), !.
bridging(drs(D1,C1),[drs(D2,C2)|K],[drs(D2,C2)|K],M) :- is_a*(drs(D1,C1),drs(.,C3)),
    m_suitable(C3,C2,[],M), !.

bridging(drs(.,C1),[drs(D2,C2)|K],[drs(D2,C)|K],M) :- imp_o(drs(D2,C2),drs(D3,C3)),
    append(C2,[drs(D3,C3)],C),
    m_suitable(C1,C3,[],M), !.

bridging(drs(D1,C1),[drs(D2,C2)|K],[drs(D2,C)|K],M) :- is_a*(drs(D1,C1),drs(.,C4)),
    imp_o(drs(D2,C2),drs(D6,C6)),
    m_suitable(C4,C6,[],M),
    append(C2,[drs(D6,C6)],C), !.

accomodate(drs(.,.),K,K,[]).

renameDRS(drs(D,C),drs(D2,R)) :- bind(D,B), subs(D,B,D2), subDRS(C,B,R).

subDRS([],-,[]).

subDRS([K1|C],B,[K2|R]) :- K1 =.. [P|A],
    subs(A,B,S),
    K2 =.. [P|S],
    subDRS(C,B,R), !.

subs([],-,[]).
subs([X|R],B,[Y|E]) :- member([X,Y],B), subs(R,B,E), !.
subs([X|R],B,[X|E]) :- subs(R,B,E).

bind([X|R],[[X,Id]|E]) :- next_id(Id), bind(R,E).
bind([],[]).

inverse([],[]).

inverse([E|R],Res) :- inverse(R,Inv), append(Inv,[E],Res).
```

B.3 reasoning.pl

```
is_a*(drs(D1,C1),drs(D2,C2)) :- isra(drs(D1,C3),drs(D2,C2)),
    match(C3,C1).
```

```

imp_o(drs(D1,C1),drs(D2,C2)) :- imp(drs(D1,C3),drs(D2,C2)),
                                match(C3,C1).

isra(DRS1,DRS2) :- is_a(DRS1,DRS2).

isra(drs(D1,C1),drs(D3,C3)) :- is_a(drs(D1,C1),drs(_,C2)),
                                isra(drs(_,C2),drs(D3,C3)).

imp(DRS1,DRS2) :- imply(DRS1,DRS2).

imp(DRS1,DRS2) :- imply(drs(_,C3),DRS2),
                       isra(DRS1,drs(_,C3)).

imply(DRS1,DRS2) :- def(DRS1,DRS2), gmp(DRS1,DRS2).

imply(drs(D,C),drs(D2,C4)) :- member(event(E1),C),
                                renameDRS(drs(D,C),drs(D2,C2)),
                                member(event(E2),C2),
                                member(X=[Y,Z|R],C2,C3),
                                append(L1,_,[Y,Z|R]),
                                append(_,[Q],L1),
                                append(C3,[X=Q,subevent(E2,E1)],C4).

imply(drs(D,C),drs(D2,C4)) :- member(event(E1),C,_),
                                renameDRS(drs(D,C),drs(D2,C2)),
                                member(event(E2),C2),
                                member(X=[Y,Z|R],C2,C3),
                                append(L1,L2,[Y,Z|R]),
                                append(_,[Q|V],L1),
                                append(_,L3,L2),
                                append(_,[U|W],L3),
                                append([Q|V],[U|W],L),
                                append(C3,[X=L,subevent(E2,E1)],C4).

match([],_).
match([E1|R1],R) :- member(E1,R,R2), match(R1,R2).

```

B.4 gmp.pl

```

gmp(drs(_,_),drs(D2,_)) :- new(D2).

new([]).
new([D|R]) :- next_id(D), new(R).

next_id(Id) :- flag(counter,Value,Value+1),

```

```
name(Value,String),  
append("id",String,String2), name(Id,String2).
```

C The Ontology O_{Bio}

This section contains the ontology O_{Bio} which has been developed as described in section 5. For each concept defined within the ontology¹⁸ its DRT representation as well as a natural language description of its intended meaning is provided. For the axioms representing taxonomic relations as well as definitions of concepts the DRT representation is given. Each axiom is preceded by a short comment on its intended meaning.

C.1 The Concepts in C_{Bio}

Definition 22 (Interaction between biochemical objects (bo))

A biochemical object $BO1$ interacts with a biochemical object $BO2$.

$E1, BO1, BO2$
$interaction(E1, BO1, BO2)$ $event(E1)$ $bo(BO1)$ $bo(BO2)$

Definition 23 (Binding between biochemical objects)

A biochemical object $BO1$ binds a biochemical object $BO2$.

$E1, BO1, BO2$
$binding(E1, BO1, BO2)$ $event(E1)$ $bo(BO1)$ $bo(BO2)$

Definition 24 (Binding between a protein and a biochemical object)

A protein P binds a biochemical object BO .

$E1, P, BO$
$binding(E1, P, BO)$ $event(E1)$ $protein(P)$ $bo(BO)$

¹⁸Concepts denoting special biochemical processes such as *cell growth, differentiation, neuronal transmission* as well as biochemical objects such as *proteins, chemical elements, protein domains* etc., which are typically represented by unary predicates are not given.

Definition 25 (Protein-protein interactions)

A protein $P1$ interacts with a protein $P2$.

$E1, P1, P2$
$interaction(E1, P1, P2)$ $event(E1)$ $protein(P1)$ $protein(P2)$

Definition 26 (Interaction between a protein and a biochemical object)

A protein P interacts with a biochemical object BO .

$E1, P, BO$
$interaction(E1, P, BO)$ $event(E1)$ $protein(P)$ $bo(BO)$

Definition 27 (Protein-source binding)

The domain $D1$ of a protein $P1$ binds a source S within an organism O .

$E1, P1, D1, S, O$
$binding_source(E1, P1, D1, S, O)$ $event(E1)$ $protein(P1)$ $protein_domain(D1)$ $subunit(D1, P1)$ $source(S)$ $organism(O)$ $subunit(S, O)$

Definition 28 (Protein-protein binding)

The domain $D1$ at location $L1$ of a protein $P1$ binds the domain $D2$ of a protein $P2$ at location $L2$ together with a protein $P3$ in presence of a substance $S1$ and in absence of a substance $S2$. The binding is characterized by an affinity A , a specificity S and the type of the binding T .

<i>E1, P1, D1, L1, P2, D2, L2, P3, S1, S2, A, S, T</i>
<i>binding(E1, P1, D1, L1, P2, D2, L2, P3, S1, S2, A, S, T)</i> <i>event(E1)</i> <i>protein(P1)</i> <i>protein_domain(D1)</i> <i>location(L1)</i> <i>subunit(D1, P1)</i> <i>protein(P2)</i> <i>protein_domain(D2)</i> <i>location(L2)</i> <i>subunit(D2, P2)</i> <i>protein(P3)</i> <i>substance(S1)</i> <i>substance(S2)</i> <i>affinity(A)</i> <i>specificity(S)</i> <i>type(T)</i> <i>binding_domain(E1, P1, D1)</i> <i>binding_domain(E1, P2, D2)</i> <i>binding_in_presence(E1, S1)</i> <i>binding_in_absence(E1, S2)</i> <i>cofactor(E1, P3)</i> <i>binding_affinity(E1, A)</i> <i>binding_specificity(E1, S)</i> <i>binding_type(E1, T)</i>

Definition 29 (Protein-protein association)

A protein *P1* associates with a protein *P2*.

<i>E1, P1, P2</i>
<i>association(E1, P1, P2)</i> <i>event(E1)</i> <i>protein(P1)</i> <i>protein(P2)</i>

Definition 30 (Cleavage)

A protein *P1* cleaves a protein *P2*.

$E1, P1, P2$
<i>cleavage</i> ($E1, P1, P2$) <i>event</i> ($E1$) <i>protein</i> ($P1$) <i>protein</i> ($P2$)

Definition 31 (Protein-DNA binding)

The domain $D1$ of a protein $P1$ binds a DNA molecule at the sequence $S1$ of an element E of a domain $D2$ with an affinity A and together with a protein $P2$.

$E1, P1, D1, DNA, D2, E, S1, A, P2$
<i>binding_dna</i> ($E1, P1, D1, DNA, D2, E, S1, A, P2$) <i>event</i> ($E1$) <i>protein</i> ($P1$) <i>protein_domain</i> ($D1$) <i>subunit</i> ($D1, P1$) <i>dna</i> (DNA) <i>dna_domain</i> ($D2$) <i>subunit</i> ($D2, DNA$) <i>dna_element</i> (E) <i>subunit</i> ($E, D2$) <i>dna_sequence</i> ($S1$) <i>affinity</i> (A) <i>protein</i> ($P2$) <i>binding_domain</i> ($E1, P1, D1$) <i>binding_domain</i> ($E1, DNA, D2$) <i>cofactor</i> ($E1, P2$) <i>binding_affinity</i> ($E1, A$)

Definition 32 (Protein-RNA binding)

The domain $D1$ of a protein $P1$ binds a RNA molecule at the sequence $S1$ of a domain $D2$ with an affinity A and together with a protein $P2$.

<i>E1, P1, D1, RNA, D2, S1, A, P2</i>
<i>binding_rna(E1, P1, D1, RNA, D2, S1, A, P2)</i> <i>event(E1)</i> <i>protein(P1)</i> <i>protein_domain(D1)</i> <i>subunit(D1, P1)</i> <i>rna(RNA)</i> <i>rna_domain(D2)</i> <i>rna_sequence(S1)</i> <i>subunit(D2, RNA)</i> <i>affinity(A)</i> <i>protein(P2)</i> <i>binding_domain(E1, P1, D1)</i> <i>binding_domain(E1, RNA, D2)</i> <i>cofactor(E1, P2)</i> <i>binding_affinity(E1, A)</i>

Definition 33 (Protein-nucleic-acid binding)

The domain *D1* of a protein *P1* binds a nucleic acid *NA* at the sequence *S1*.

<i>E1, P1, D1, NA, S1</i>
<i>binding_na(E1, P1, D1, NA, S)</i> <i>event(E1)</i> <i>protein(P1)</i> <i>protein_domain(D1)</i> <i>subunit(D1, P1)</i> <i>nucleic_acid(NA)</i> <i>na_sequence(S)</i> <i>binding_domain(E1, P1, D1)</i>

Definition 34 (Protein-gene binding)

The domain *D1* of a protein *P1* binds the sequence *S1* within an element *E* of the domain *D2* of a gene *G* with an affinity *A* and together with protein *P2*.

$E1, P1, D1, G, D2, E, S1, A, P2$
<i>binding_gene</i> ($E1, P1, D1, G, D2, E, S1, A, P2$) <i>event</i> ($E1$) <i>protein</i> ($P1$) <i>protein_domain</i> ($D1$) <i>subunit</i> ($D1, P1$) <i>gene</i> (G) <i>gene_domain</i> ($D2$) <i>subunit</i> ($D2, G$) <i>dna_element</i> (E) <i>subunit</i> ($E, D2$) <i>dna_sequence</i> ($S1$) <i>affinity</i> (A) <i>protein</i> ($P2$) <i>binding_domain</i> ($E1, P1, D1$) <i>binding_domain</i> ($E1, G, D2$) <i>binding_element</i> ($E1, D2, E$) <i>cofactor</i> ($E1, P2$) <i>binding_affinity</i> ($E1, A$)

Definition 35 (Binding between a protein and a chemical element)
The domain D of a protein P binds a chemical element C . The binding is characterized by an affinity A and a type T .

$E1, P, D, C, T, A$
<i>binding_chem</i> ($E1, P, D, C, T, A$) <i>event</i> ($E1$) <i>protein</i> (P) <i>protein_domain</i> (D) <i>subunit</i> (D, P) <i>chemical_element</i> (C) <i>type</i> (T) <i>affinity</i> (A) <i>binding_domain</i> ($E1, P, D$) <i>binding_type</i> ($E1, T$) <i>binding_affinity</i> ($E1, A$)

Definition 36 (Binding between a protein and N molecules of a chemical element)
The domain D of a protein P binds N molecules of a chemical element C . The binding is characterized by an affinity A and a type T .

$E1, P, D, C, N, T, A$
<i>binding_chem</i> ($E1, P, D, C, N, T, A$) <i>event</i> ($E1$) <i>protein</i> (P) <i>protein_domain</i> (D) <i>subunit</i> (D, P) <i>chemical_element</i> (C) <i>number_of_molecules</i> (N) <i>type</i> (T) <i>affinity</i> (A) <i>binding_domain</i> ($E1, P, D$) <i>binding_n</i> ($E1, N$) <i>binding_type</i> ($E1, T$) <i>binding_affinity</i> ($E1, A$)

Definition 37 (Binding between a protein and M moles of a chemical element)

The domain D of a protein P binds M moles of a chemical element C . The binding is characterized by an affinity A and a type T .

$E1, P, D, C, M, T, A$
<i>binding_chem</i> ($E1, P, D, C, M, T, A$) <i>event</i> ($E1$) <i>protein</i> (P) <i>protein_domain</i> (D) <i>subunit</i> (D, P) <i>chemical_element</i> (C) <i>number_of_moles</i> (M) <i>type</i> (T) <i>affinity</i> (A) <i>binding_domain</i> ($E1, P, D$) <i>binding_m</i> ($E1, M$) <i>binding_type</i> ($E1, T$) <i>binding_affinity</i> ($E1, A$)

Definition 38 (Protein-ion binding)

The domain D of a protein P binds N molecules of a ion C . The binding is characterized by an affinity A and a type T .

$E1, P, D, C, N, T, A$
$binding_ion(E1, P, D, C, N, T, A)$ $event(E1)$ $protein(P)$ $protein_domain(D)$ $subunit(D, P)$ $ion(C)$ $number_of_molecules(N)$ $type(T)$ $affinity(A)$ $binding_domain(E1, P, D)$ $binding_n(E1, N)$ $binding_type(E1, T)$ $binding_affinity(E1, A)$

Definition 39 (Activation of a gene)

A protein $P1$ activates a gene G .

$E1, P1, G$
$activation(E1, P1, G)$ $event(E1)$ $protein(P1)$ $gene(G)$

Definition 40 (Repression of a gene)

A protein $P1$ represses a gene G .

$E1, P1, G$
$repression(E1, P1, G)$ $event(E1)$ $protein(P1)$ $gene(G)$

Definition 41 (Inactivation of a gene)

A protein $P1$ inactivates a gene G .

$E1, P1, G$
<i>inactivation</i> ($E1, P1, G$) <i>event</i> ($E1$) <i>protein</i> ($P1$) <i>gene</i> (G)

Definition 42 (Inhibition of a gene)

A protein $P1$ inhibits a gene G .

$E1, P1, G$
<i>inhibition</i> ($E1, P1, G$) <i>event</i> ($E1$) <i>protein</i> ($P1$) <i>gene</i> (G)

Definition 43 (Control/Regulation of a biochemical process)

A protein P control/regulates a biochemical process $E2$.

$E1, P, E2$
<i>control_regulation</i> ($E1, P, E2$) <i>event</i> ($E1$) <i>protein</i> (P) <i>biochemical_process</i> ($E2$)

Definition 44 (Acceleration of a biochemical process)

A protein P accelerates a biochemical process $E2$.

$E1, P, E2$
<i>acceleration</i> ($E1, P, E2$) <i>event</i> ($E1$) <i>protein</i> (P) <i>biochemical_process</i> ($E2$)

Definition 45 (Mediation of a biochemical process)

A protein P mediates a biochemical process $E2$.

$E1, P, E2$
$mediation(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 46 (Induction of a biochemical process)

A protein P induces a biochemical process $E2$.

$E1, P, E2$
$induction(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 47 (Catalyzation of a biochemical process)

A protein P catalyzes a biochemical process $E2$.

$E1, P, E2$
$catalyzation(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 48 (Inhibition of a biochemical process)

A protein P inhibits a biochemical process $E2$.

$E1, P, E2$
$inhibition(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 49 (Stimulation of a biochemical process)

A protein P stimulates a biochemical process $E2$.

$E1, P, E2$
$stimulation(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 50 (Enhancement of a biochemical process)

A protein P enhances a biochemical process $E2$.

$E1, P, E2$
$enhancement(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 51 (Facilitation of a biochemical process)

A protein P facilitates a biochemical process $E2$.

$E1, P, E2$
$facilitation(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 52 (Promotion of a biochemical process)

A protein P promotes a biochemical process $E2$.

$E1, P, E2$
$promotion(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 53 (Regulation of a biochemical process)

A protein P regulates a biochemical process $E2$.

$E1, P, E2$
$regulation(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 54 (Control of a biochemical process)

A protein P controls a biochemical process $E2$.

$E1, P, E2$
$control(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 55 (Activation of a biochemical process)

A protein P activates a biochemical process $E2$.

$E1, P, E2$
$activation(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 56 (Repression of a biochemical process)

A protein P represses a biochemical process $E2$.

$E1, P, E2$
$repression(E1, P, E2)$ $event(E1)$ $protein(P)$ $biochemical_process(E2)$

Definition 57 (Suppression of a biochemical process)

A protein P suppresses a biochemical process $E2$.

$E1, P, E2$
<i>suppression(E1, P, E2)</i> <i>event(E1)</i> <i>protein(P)</i> <i>biochemical_process(E2)</i>

Definition 58 (Blocking of a biochemical process)

A protein P blocks a biochemical process $E2$.

$E1, P, E2$
<i>blocking(E1, P, E2)</i> <i>event(E1)</i> <i>protein(P)</i> <i>biochemical_process(E2)</i>

Definition 59 (Prevention of a biochemical process)

A protein P prevents a biochemical process $E2$.

$E1, P, E2$
<i>prevention(E1, P, E2)</i> <i>event(E1)</i> <i>protein(P)</i> <i>biochemical_process(E2)</i>

Definition 60 (Change of location)

A protein $P1$ changes the location of a protein $P2$ from a cellular location $L1$ into a cellular location $L2$.

$E1, P1, P2, L1, L2$
<i>change_of_loc(E1, P1, P2, L1, L2)</i> <i>event(E1)</i> <i>protein(P1)</i> <i>protein(P2)</i> <i>source(L1)</i> <i>source(L2)</i> <i>from(E1, L1)</i> <i>goal(E1, L2)</i>

Definition 61 (Collection of a protein)

A protein $P1$ collects a protein $P2$ from a cellular location $L1$ into a cellular location $L2$.

$E2, P1, P2, L1, L2$
$collection(E2, P1, P2, L1, L2)$ $event(E2)$ $protein(P1)$ $protein(P2)$ $source(L1)$ $source(L2)$ $from(E2, L1)$ $goal(E2, L2)$

Definition 62 (Recruitment of a substance)

A protein $P1$ recruits a protein $P2$ from a cellular location $L1$ into a cellular location $L2$.

$E2, P1, P2, L1, L2$
$recruitment(E2, P1, P2, L1, L2)$ $event(E2)$ $protein(P1)$ $protein(P2)$ $source(L1)$ $source(L2)$ $from(E2, L1)$ $goal(E2, L2)$

Definition 63 (Translocation of a protein)

A protein $P1$ translocates a protein $P2$ from a cellular location $L1$ into a cellular location $L2$.

$E2, P1, P2, L1, L2$
$translocation(E2, P1, P2, L1, L2)$ $event(E2)$ $protein(P1)$ $protein(P2)$ $source(L1)$ $source(L2)$ $from(E2, L1)$ $goal(E2, L2)$

Definition 64 (Transport of a protein)

A protein $P1$ transports a protein $P2$ from a cellular location $L1$ into a cellular location $L2$.

$E2, P1, P2, L1, L2$
$transport(E2, P1, P2, L1, L2)$ $event(E2)$ $protein(P1)$ $protein(P2)$ $source(L1)$ $source(L2)$ $from(E2, L1)$ $goal(E2, L2)$

Definition 65 (Phosphorylation of a protein)

A protein $P1$ phosphorylates a protein $P2$.

$E2, P1, P2$
$phosphorylation(E2, P1, P2)$ $event(E2)$ $protein(P1)$ $protein(P2)$

Definition 66 (Hydrolysis of a protein)

A protein $P1$ hydrolyzes a protein $P2$.

$E2, P1, P2$
$hydrolysis(E2, P1, P2)$ $event(E2)$ $protein(P1)$ $protein(P2)$

Definition 67 (Reaction)

A protein $P1$ reacts with protein $P2$.

$E1, P1, P2$
$reaction(E1, P1, P2)$ $event(E1)$ $protein(P1)$ $protein(P2)$

Definition 68 (Modification of a protein)

A protein $P1$ modifies a protein $P2$.

$E1, P1, P2$
$modification(E1, P1, P2)$ $event(E1)$ $protein(P1)$ $protein(P2)$

Definition 69 (Inactivation of a protein)

A protein $P1$ inactivates a protein $P2$.

$E1, P1, P2$
$inactivation(E1, P1, P2)$ $event(E1)$ $protein(P1)$ $protein(P2)$

Definition 70 (Activation of a protein)

A protein $P1$ activates a protein $P2$.

$E1, P1, P2$
$activation(E1, P1, P2)$ $event(E1)$ $protein(P1)$ $protein(P2)$

Definition 71 (Modulation of a protein)

A protein $P1$ modulates a protein $P2$.

$E1, P1, P2$
$modulation(E1, P1, P2)$ $event(E1)$ $protein(P1)$ $protein(P2)$

Definition 72 (Inhibitions of a protein)

A protein $P1$ inhibits a protein $P2$.

$E1, P1, P2$
<i>inhibition</i> ($E1, P1, P2$) <i>event</i> ($E1$) <i>protein</i> ($P1$) <i>protein</i> ($P2$)

Definition 73 (Formation of a complex)

A protein $P1$ and a protein $P2$ form a complex C in state S of which they are part of.

$E1, C, S, P1, P2$
<i>formation</i> ($E1, C$) <i>event</i> ($E1$) <i>state</i> (S) <i>complex</i> (S, C) <i>protein</i> ($P1$) <i>protein</i> ($P2$) <i>part_of</i> ($P1, C$) <i>part_of</i> ($P2, C$)

Definition 74 (Protein complex)

Protein $P1$ and protein $P2$ form a complex in state S .

$S, C, P1, P2$
<i>complex</i> (S, C) <i>state</i> (S) <i>part_of</i> ($P1, C$) <i>part_of</i> ($P2, C$) <i>protein</i> ($P1$) <i>protein</i> ($P2$)

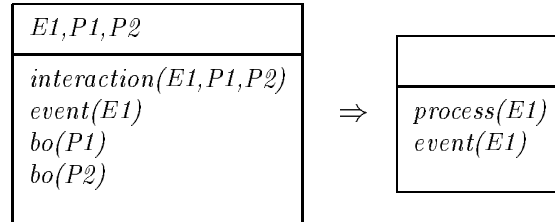
Definition 75 (Recognition of a sequence)

A protein P recognizes a specific DNA sequence.

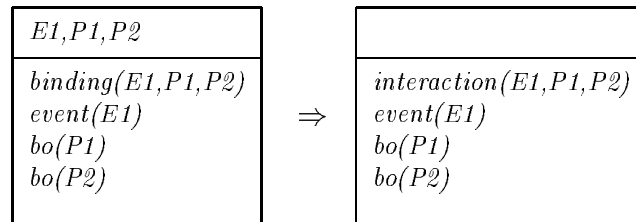
$E1, P, S$
<i>recognize</i> ($E1, P, S$) <i>event</i> ($E1$) <i>protein</i> (P) <i>dna_sequence</i> (S)

C.2 The Taxonomic Relations T_{Bio}

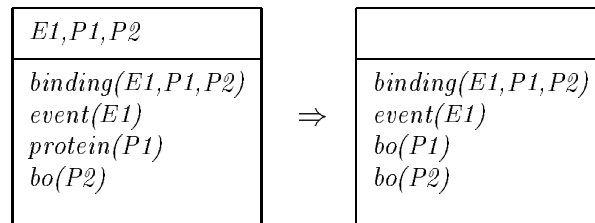
An interaction is some type of process:



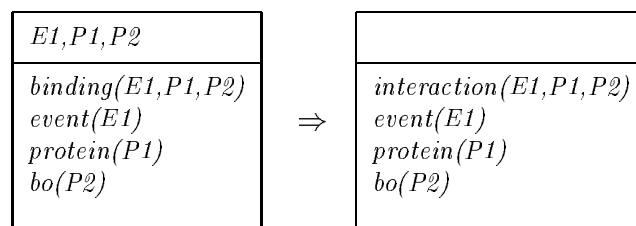
A binding between two biochemical objects is an interaction:



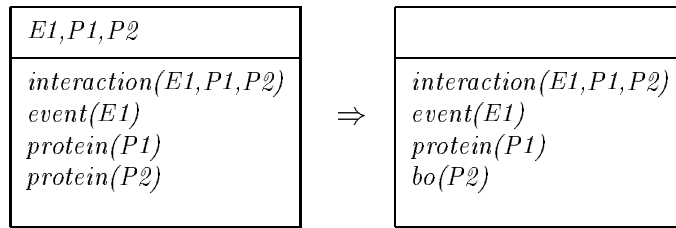
A binding between a protein and a biochemical object is a binding between two biochemical objects:



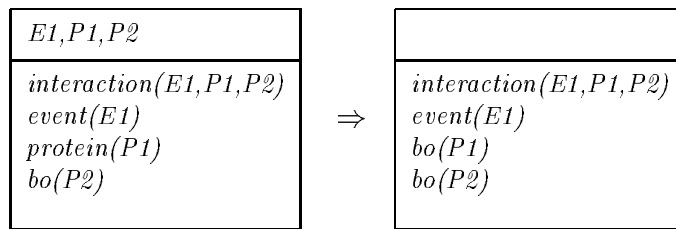
A binding between a protein and a biochemical object is an interaction between a protein and a biochemical object:



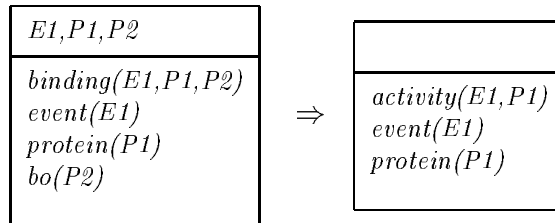
An interaction between a protein and a protein is an interaction between a protein and a biochemical object:



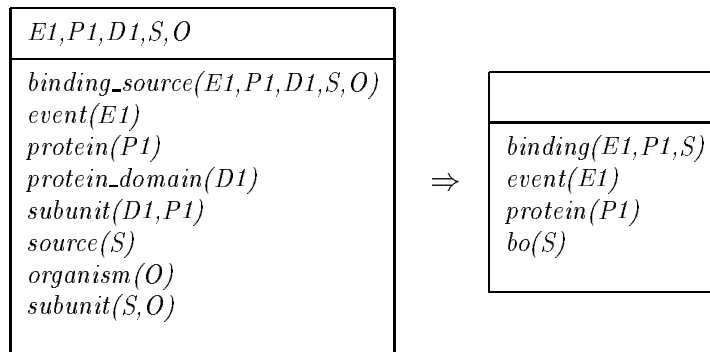
An interaction between a protein and a biochemical object is an interaction between two biochemical objects:



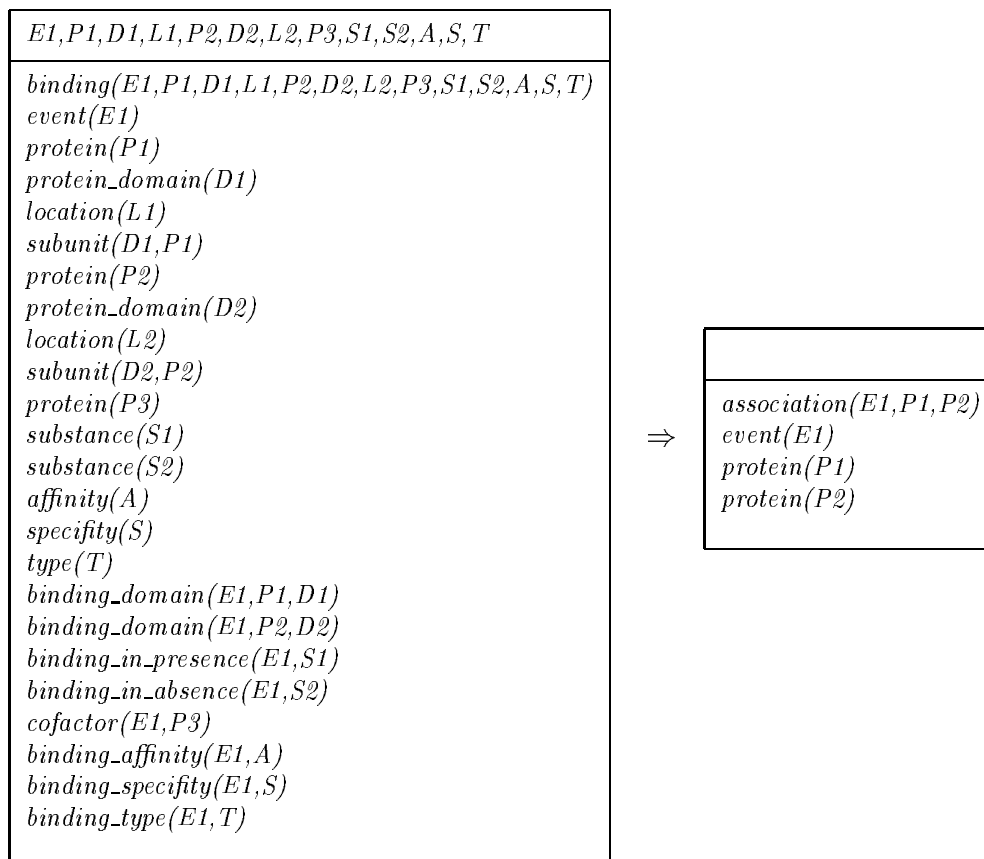
A binding between a protein and a biochemical object is an activity of the protein:



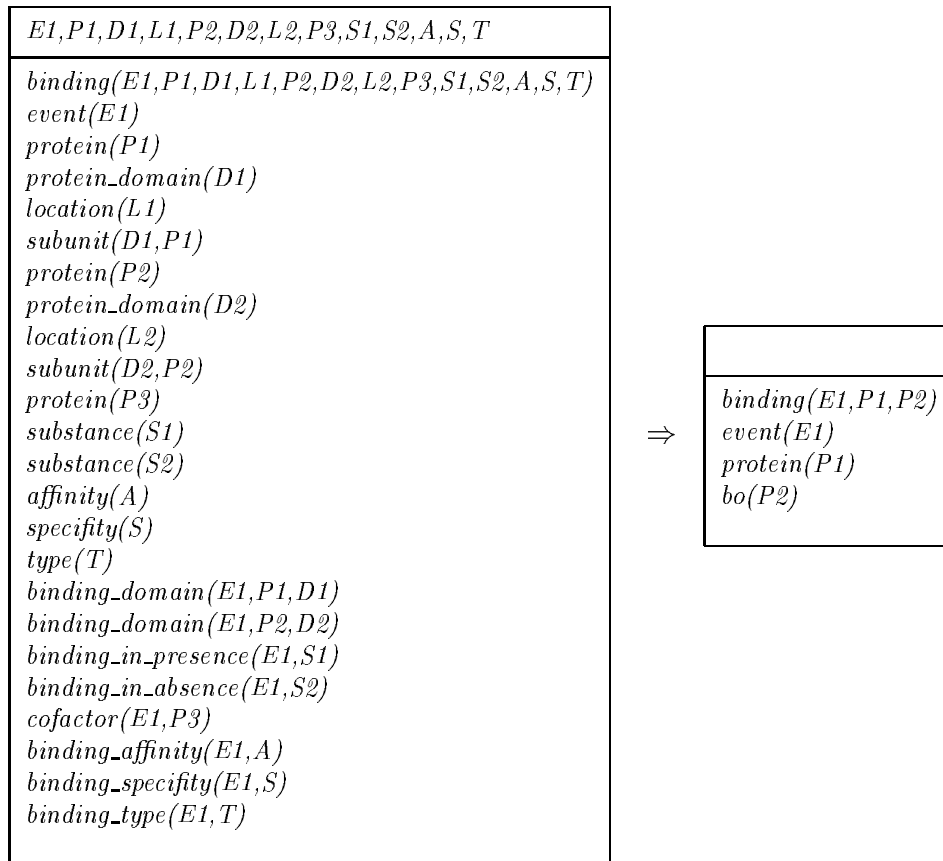
A binding between a protein and a source is a binding between a protein and a biochemical object:



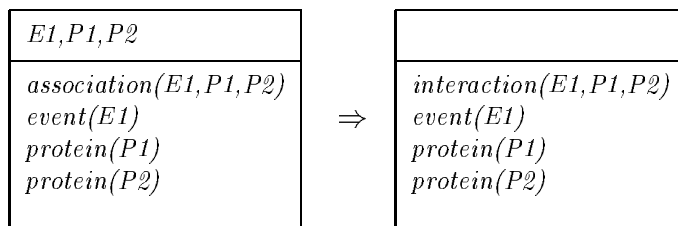
A binding between two proteins is an association:



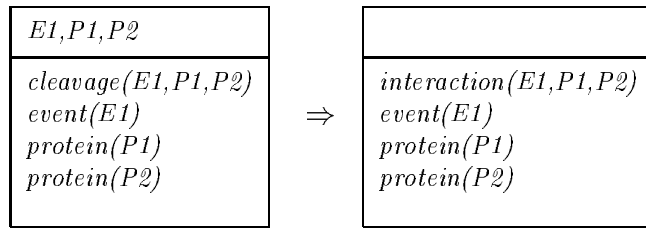
But it is also a binding between a protein and a biochemical object:



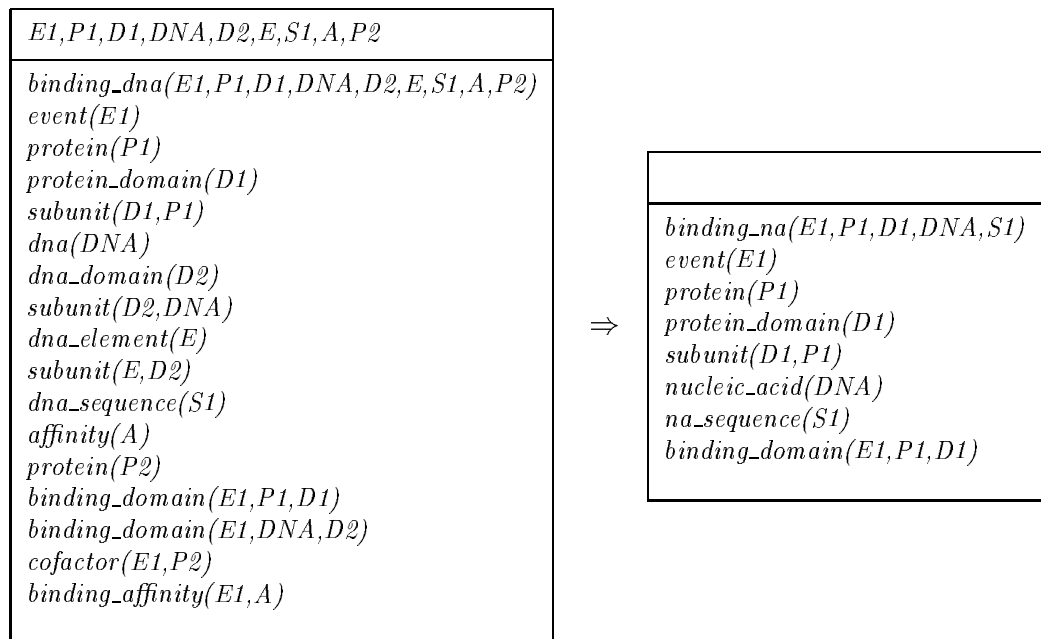
An association is an interaction between a protein and a protein:



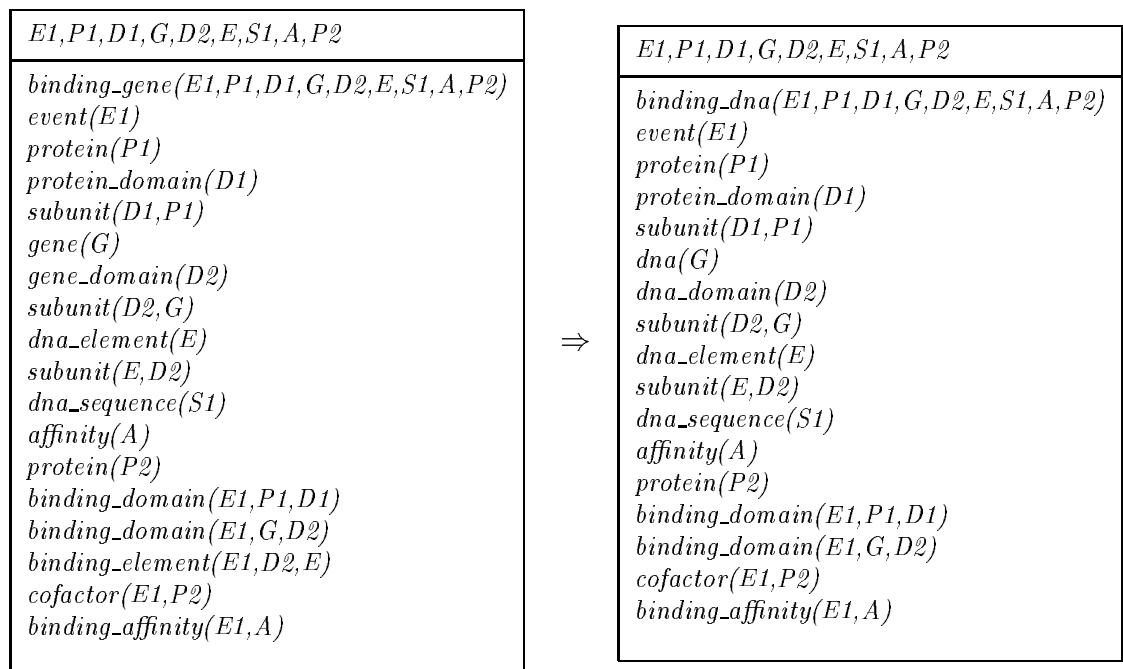
A cleavage is also an interaction between proteins:



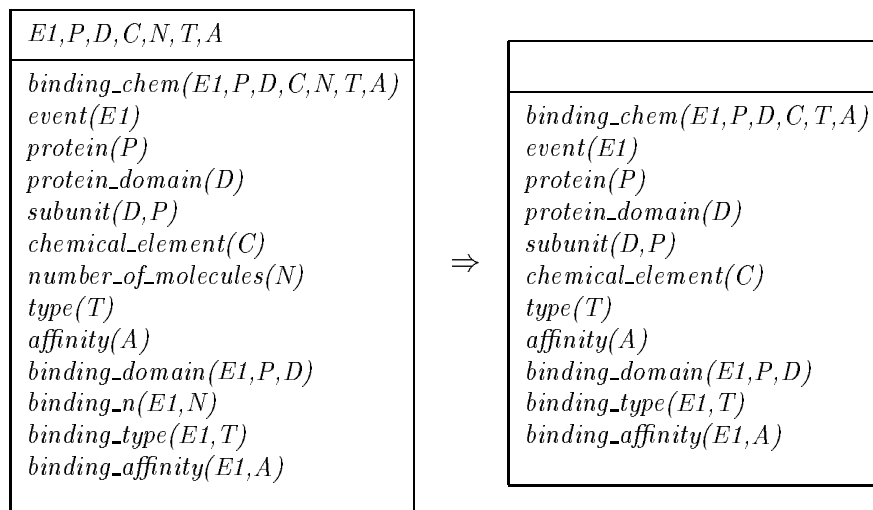
A binding between a protein and DNA is a binding between a protein and a nucleic acid:



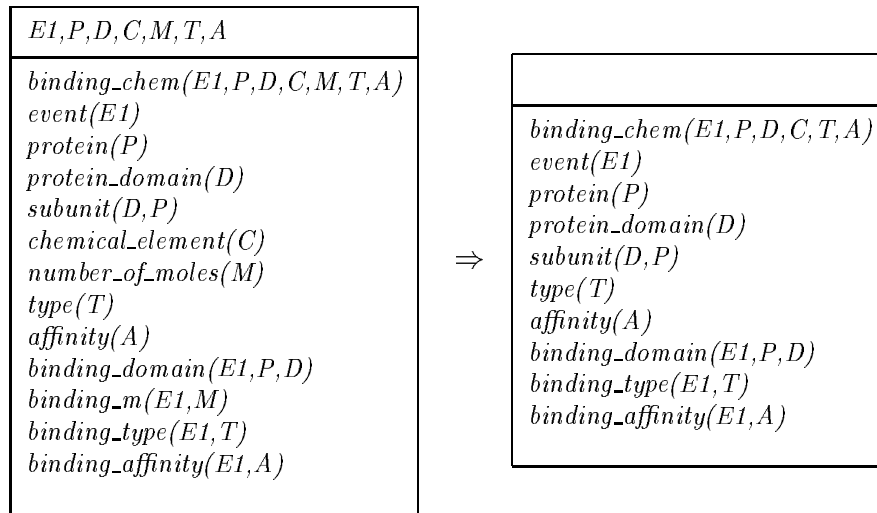
A binding between a protein and RNA is a binding between a protein and a nucleic acid:



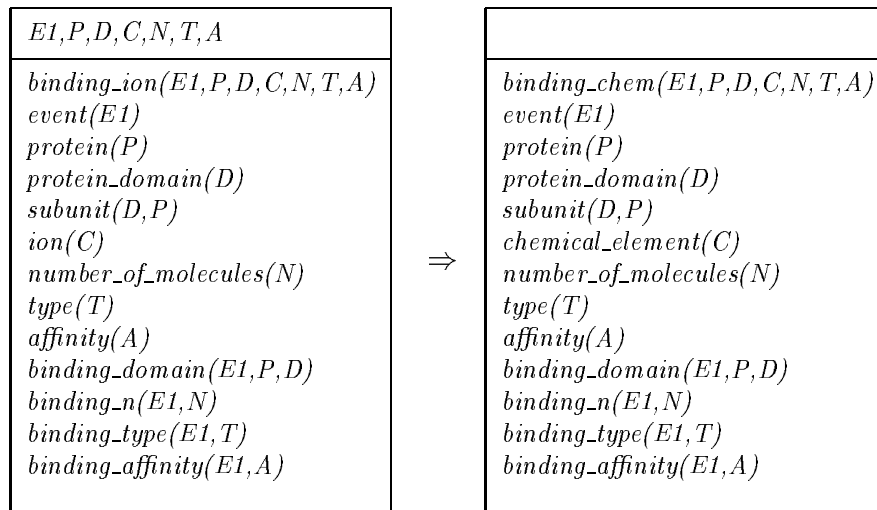
A binding between a protein and n molecules of a chemical element is a binding between a protein and a chemical element:



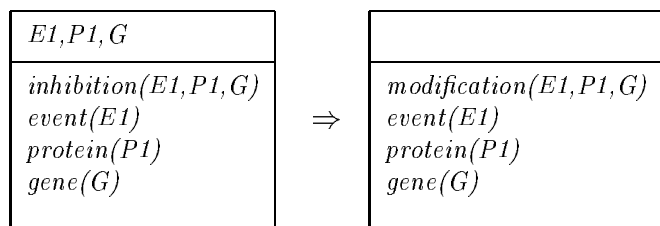
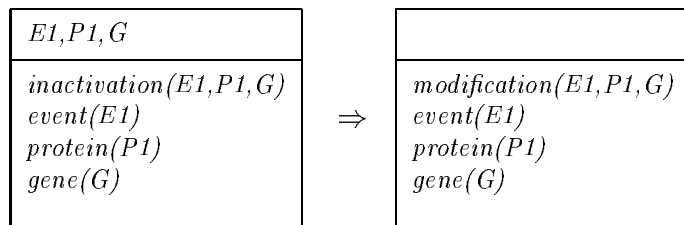
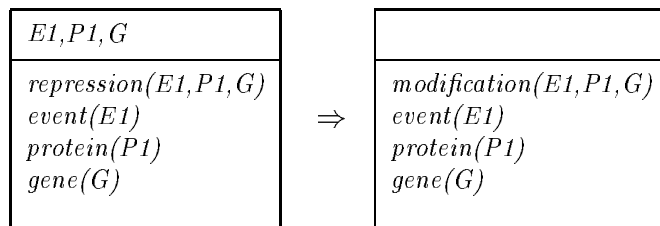
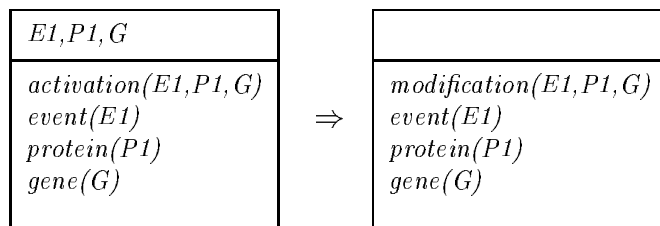
A binding between a protein and m moles of a chemical element is a binding between a protein and a chemical element:



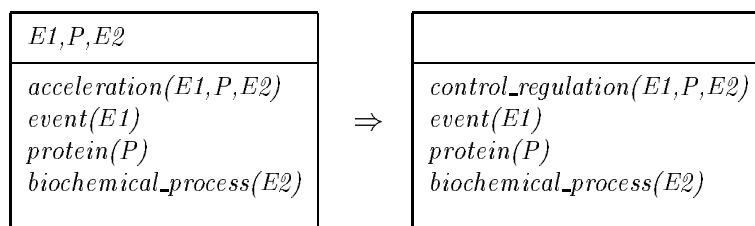
A binding between a protein and n ions is a binding between a protein and n molecules chemical element :

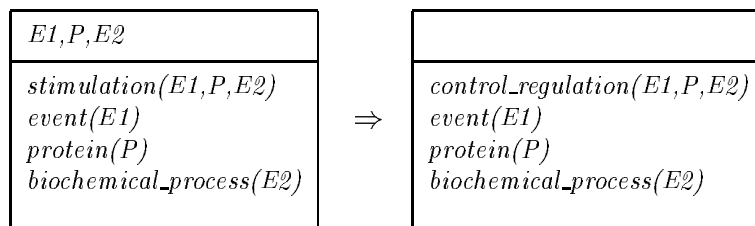
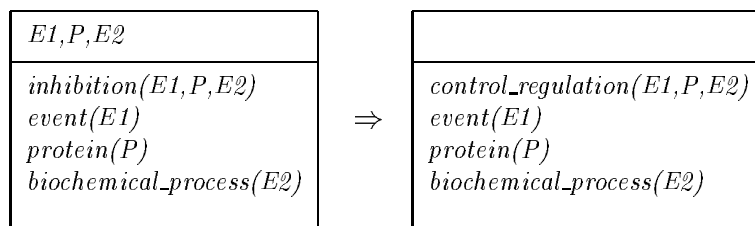
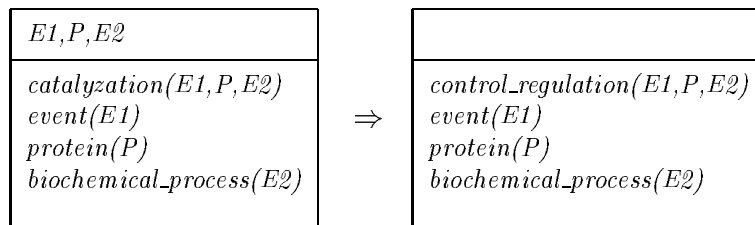
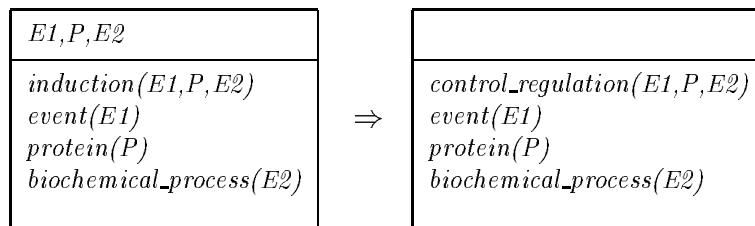
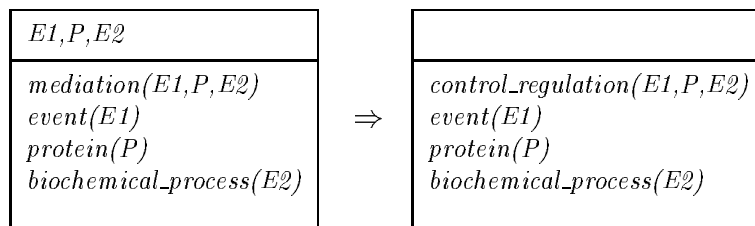


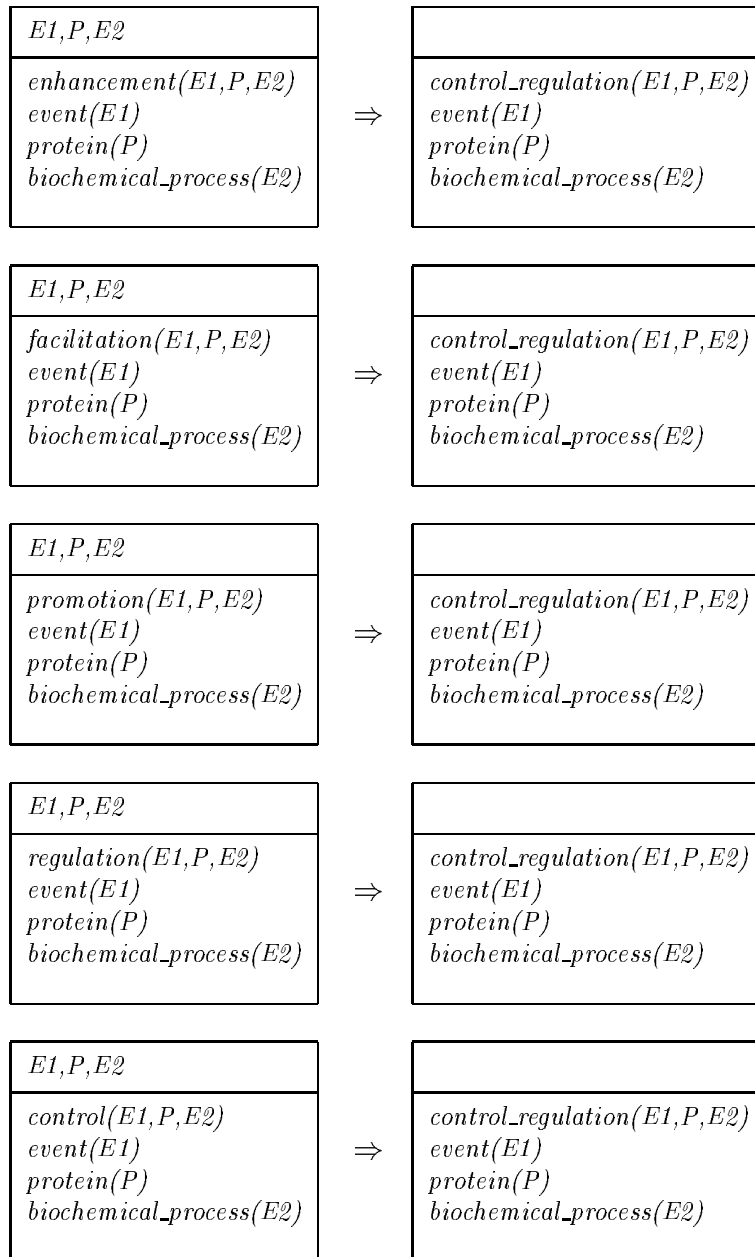
An activation / repression / inhibition / inactivation of a gene are all modification events:

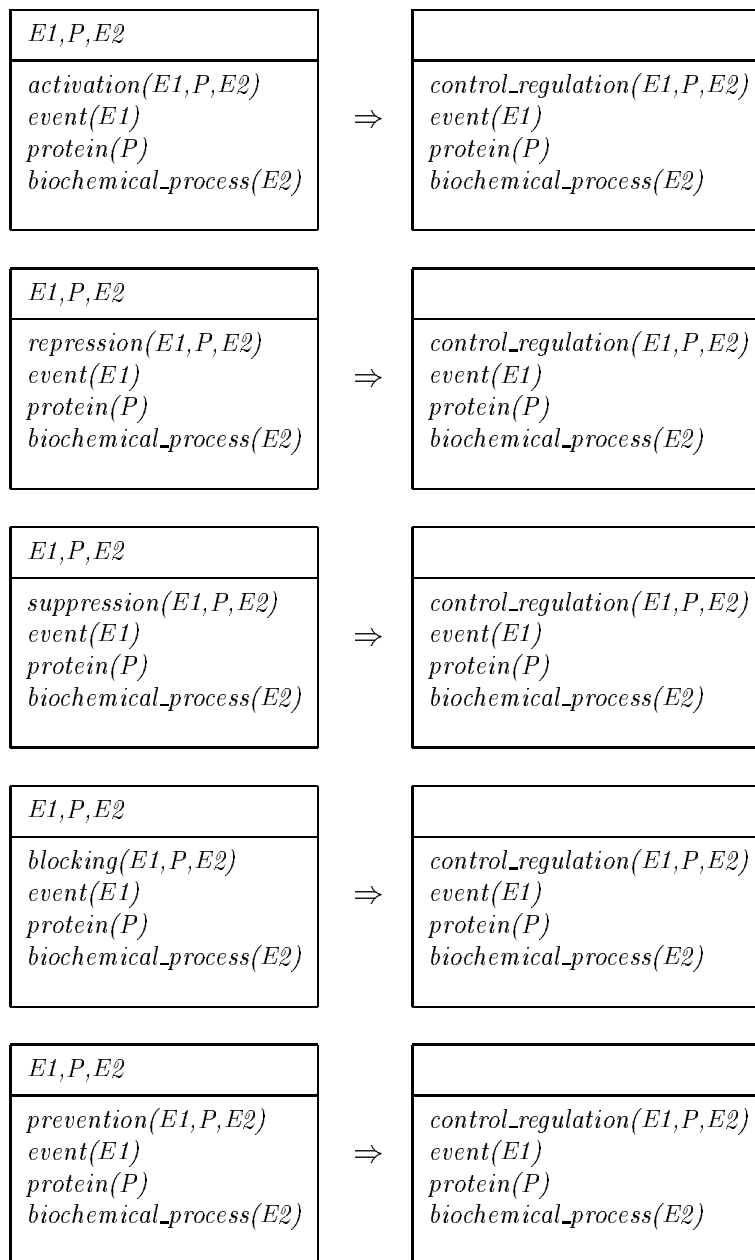


An inhibition, stimulation, repression or blocking are types of control/regulation events:

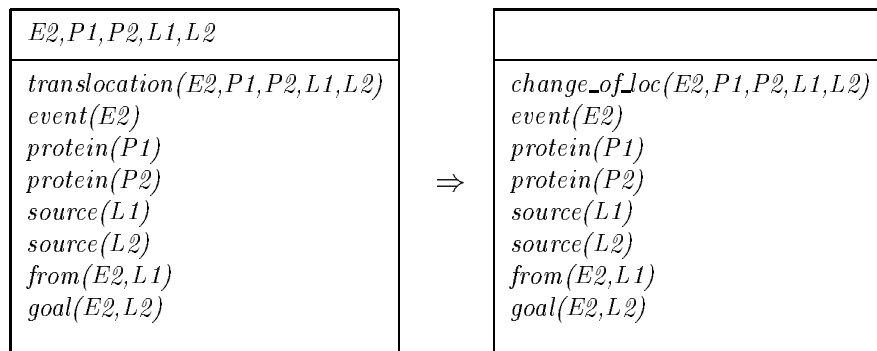
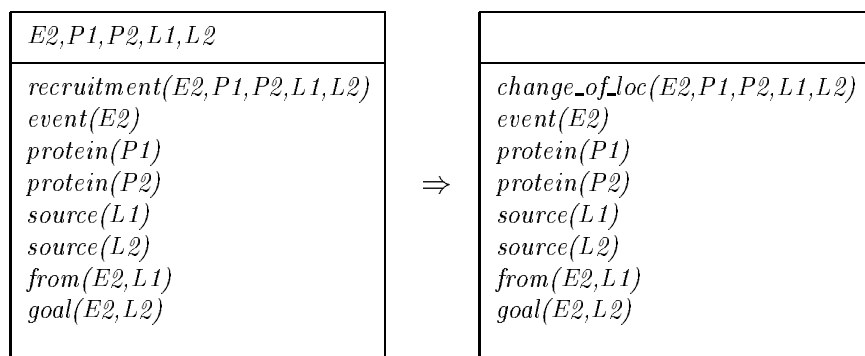
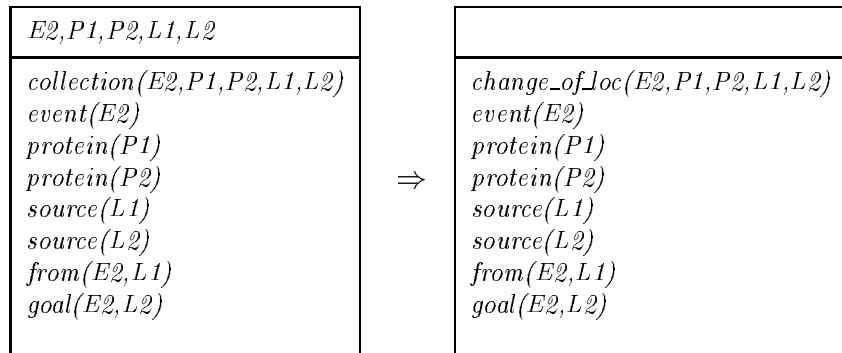


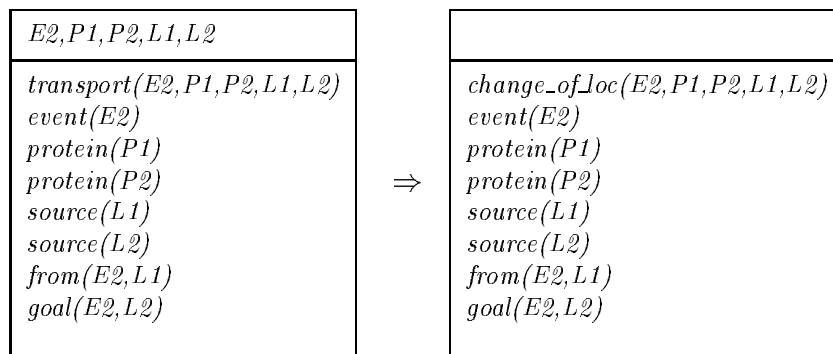




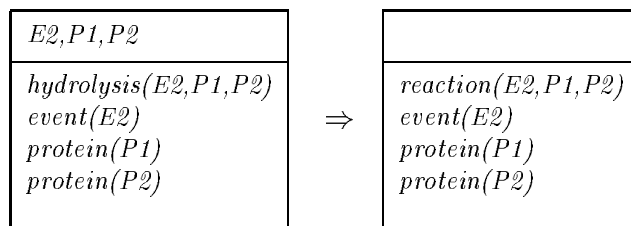
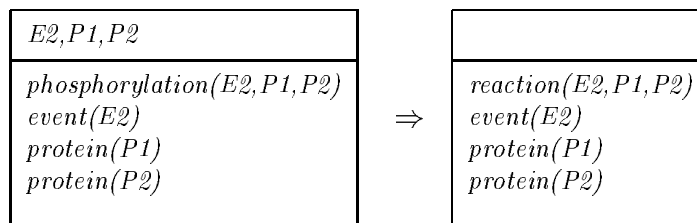


A collection, a recruitment, a translocation, a transport are all types of change_of_location events:

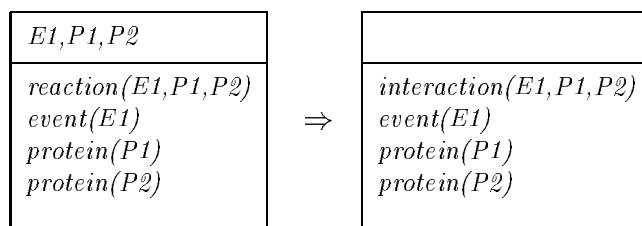




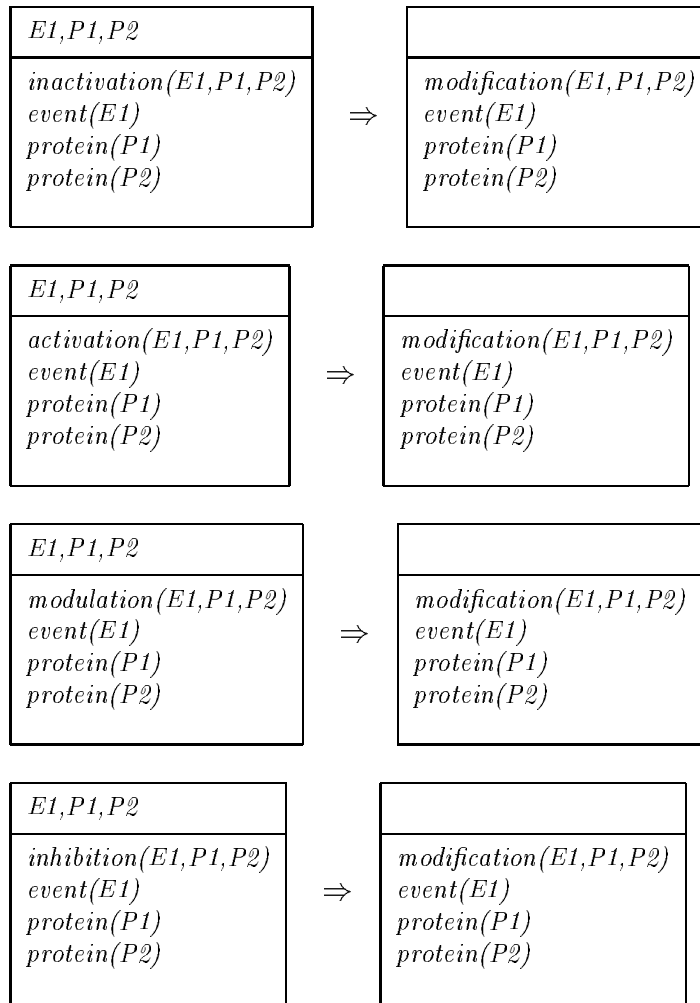
A phosphorylation is a chemical reaction:



A reaction between two proteins is some type of interaction:

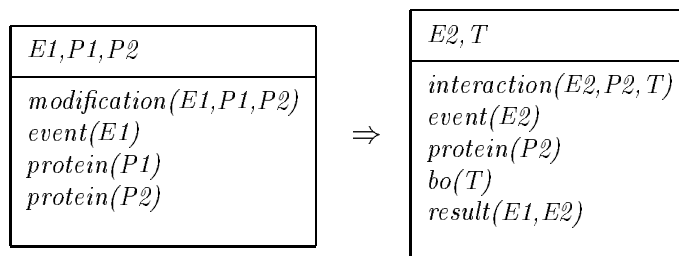


An inactivation, inhibition or activation of a protein are modification events:

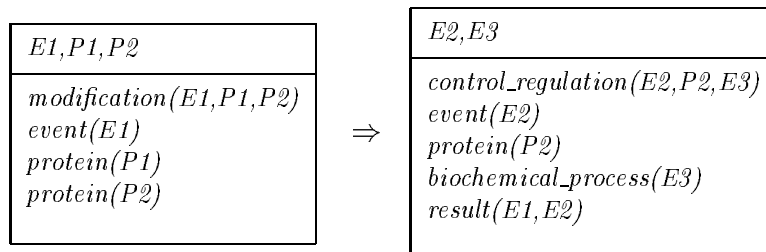


C.3 The Conceptual Definitions D_{Bio}

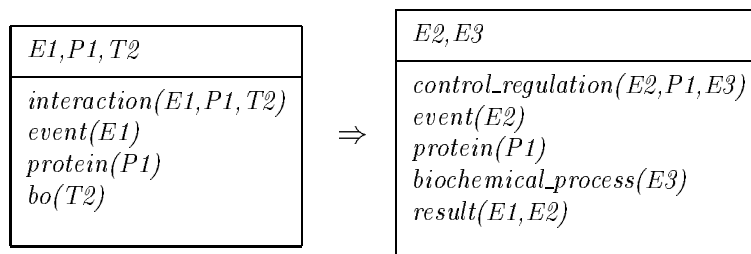
A modification of a protein leads to a further interaction of it:



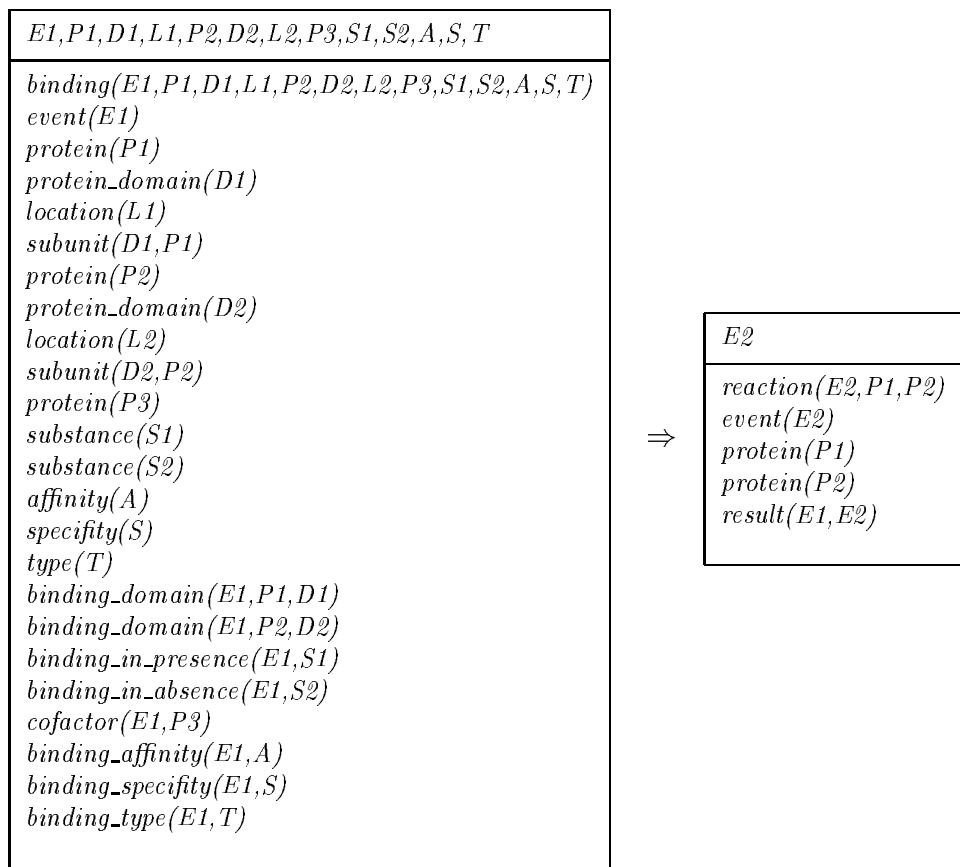
A modification of a protein leads to a control/regulation event:



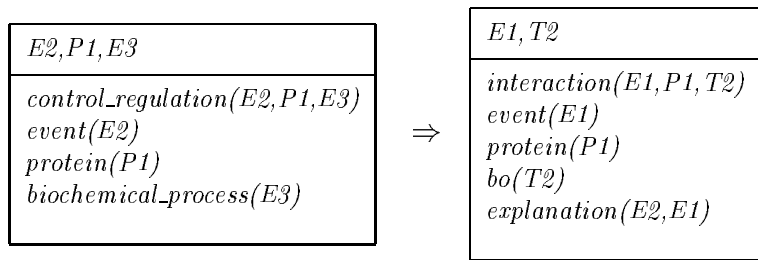
An interaction between a protein and a biochemical object normally serves to control/regulate some biochemical process:

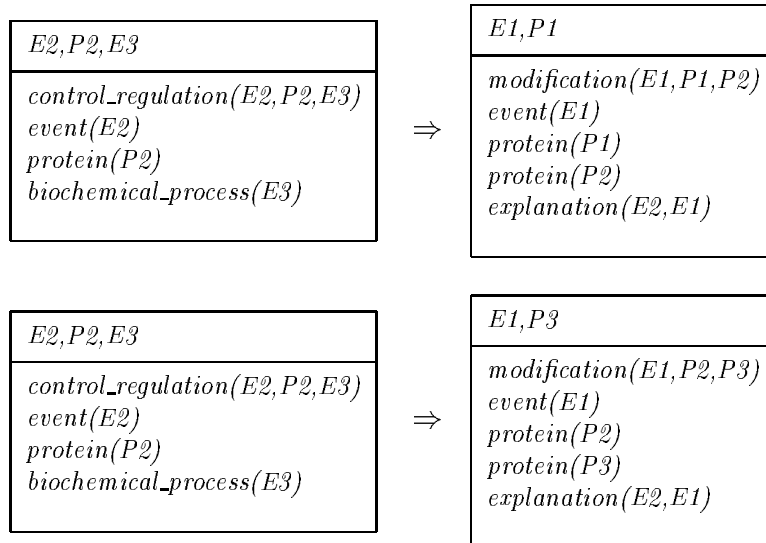


A binding between a protein and a protein normally leads to a certain reaction:

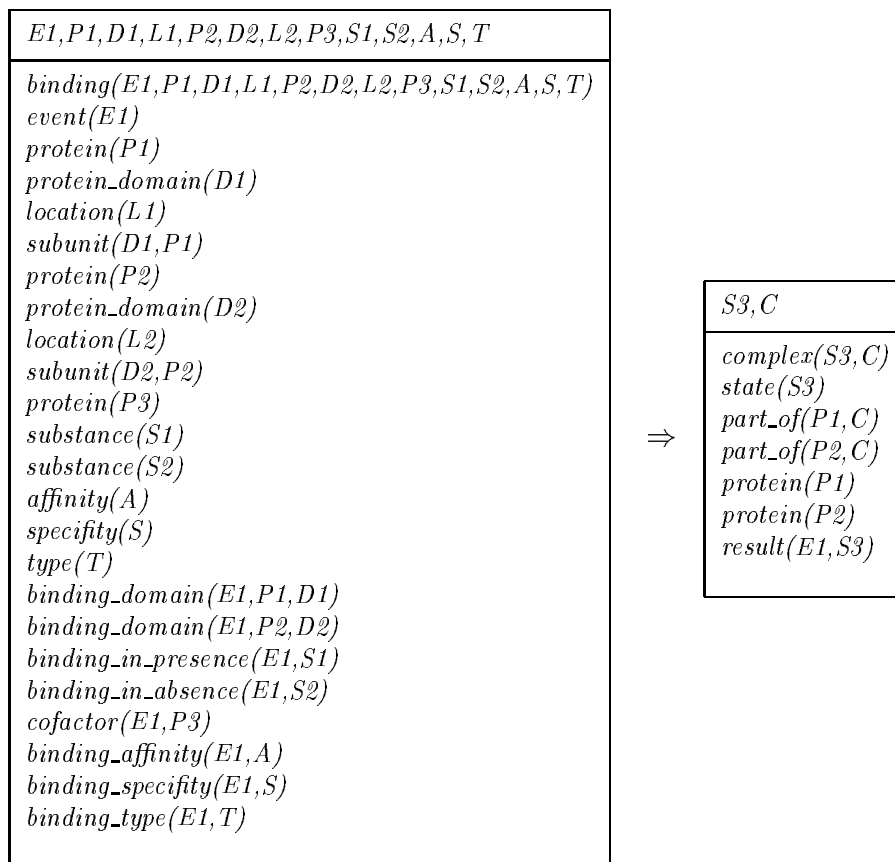


A control/regulation event presupposes a certain interaction/modification:





If two proteins bind together they form a complex as a result:



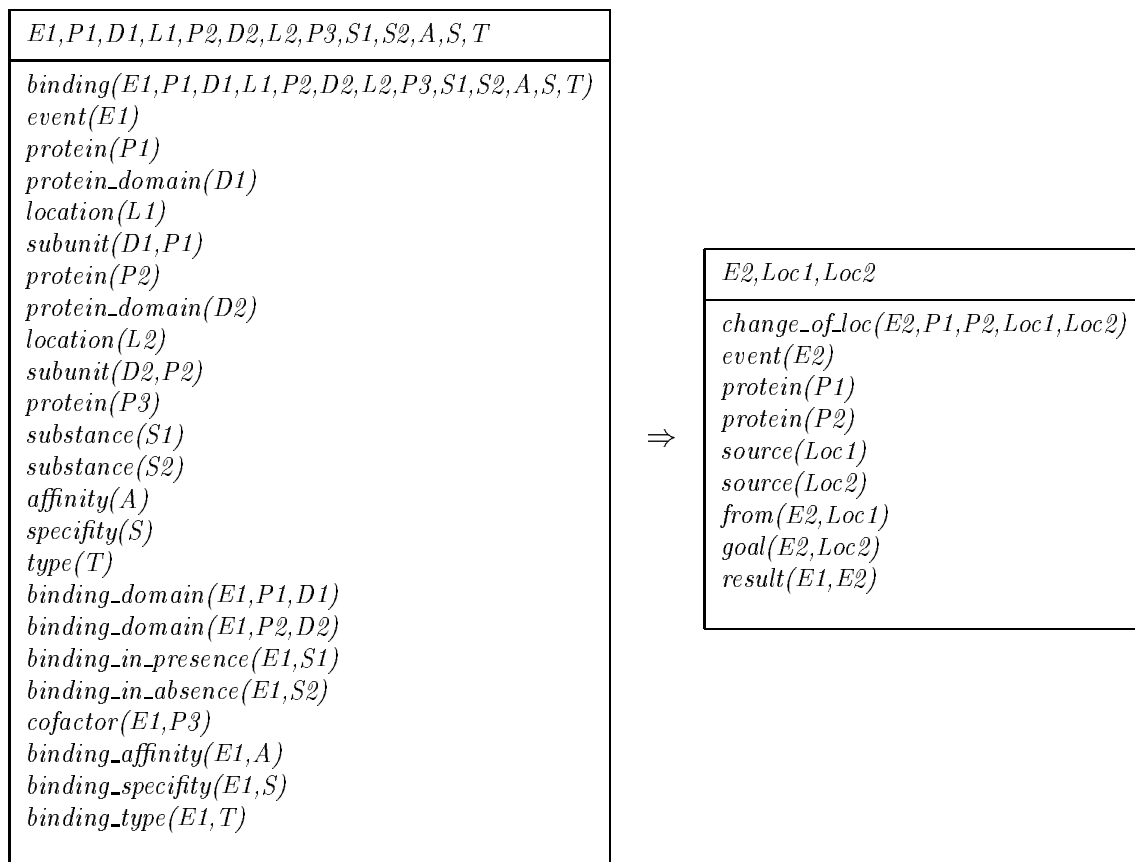
A binding of a protein to a protein leads to the formation of a complex:

<i>E1, P1, D1, L1, P2, D2, L2, P3, S1, S2, A, S, T</i>
<i>binding(E1, P1, D1, L1, P2, D2, L2, P3, S1, S2, A, S, T)</i> <i>event(E1)</i> <i>protein(P1)</i> <i>protein_domain(D1)</i> <i>location(L1)</i> <i>subunit(D1, P1)</i> <i>protein(P2)</i> <i>protein_domain(D2)</i> <i>location(L2)</i> <i>subunit(D2, P2)</i> <i>protein(P3)</i> <i>substance(S1)</i> <i>substance(S2)</i> <i>affinity(A)</i> <i>specificity(S)</i> <i>type(T)</i> <i>binding_domain(E1, P1, D1)</i> <i>binding_domain(E1, P2, D2)</i> <i>binding_in_presence(E1, S1)</i> <i>binding_in_absence(E1, S2)</i> <i>cofactor(E1, P3)</i> <i>binding_affinity(E1, A)</i> <i>binding_specificity(E1, S)</i> <i>binding_type(E1, T)</i>

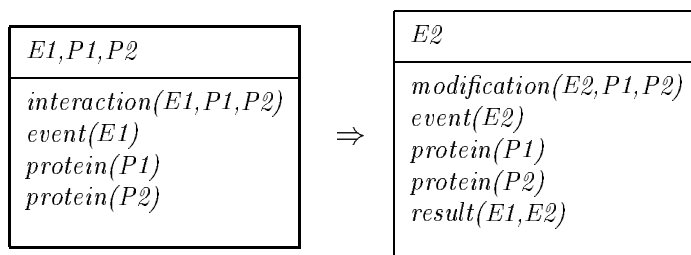
⇒

<i>E2, C, S3, P4, P5</i>
<i>formation(E2, C)</i> <i>event(E2)</i> <i>state(S3)</i> <i>complex(S3, C)</i> <i>protein(P4)</i> <i>protein(P5)</i> <i>part_of(P4, C)</i> <i>part_of(P5, C)</i> <i>result(E1, E2)</i> <i>result(S3, E2)</i>

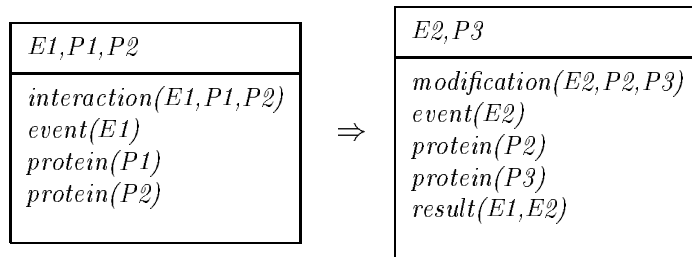
A binding of a protein by a protein can lead to a change_of_location event:



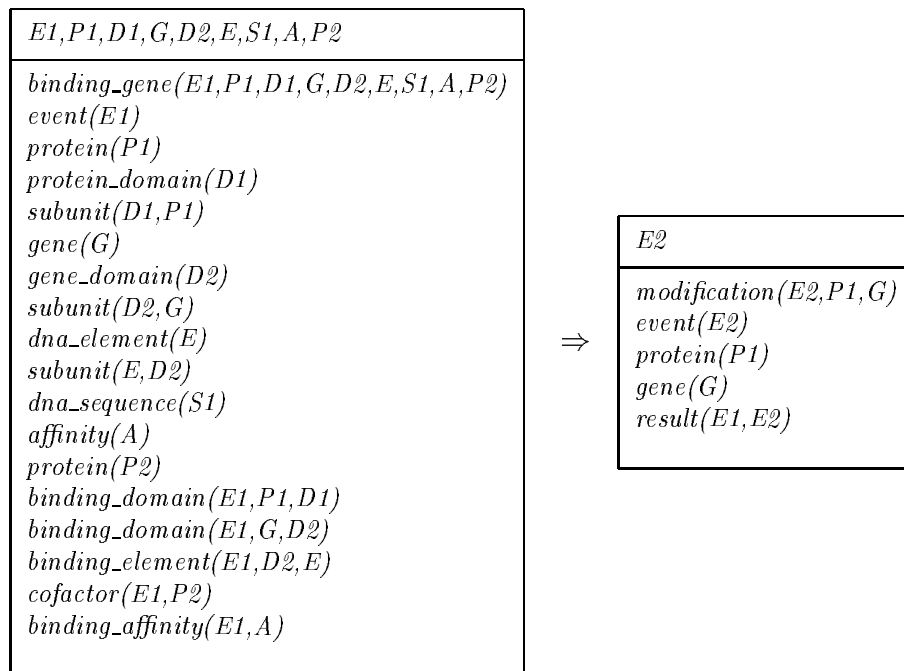
An interaction between two proteins leads to a modification of the second one by the first one:



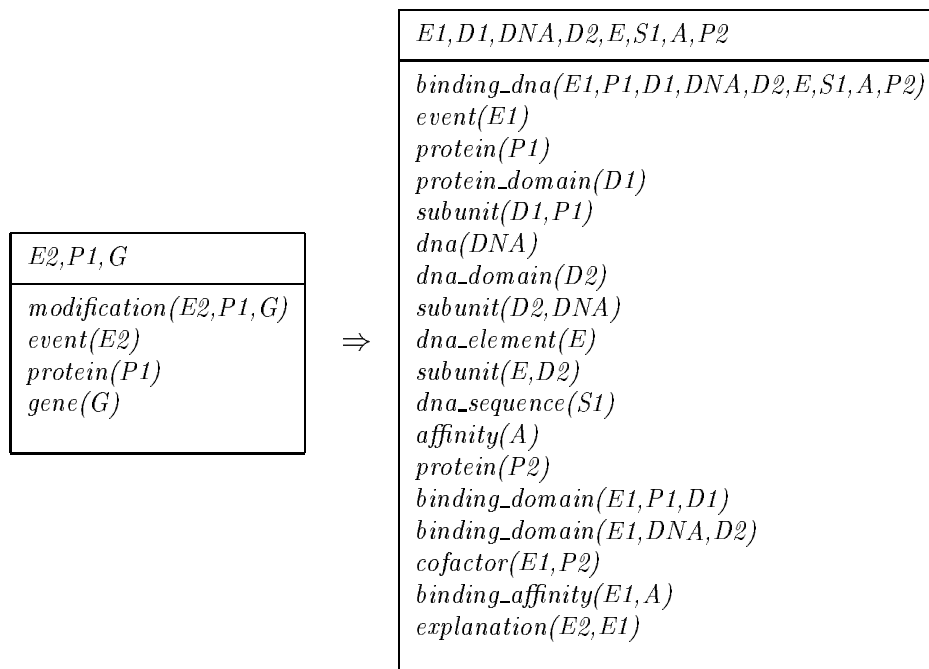
or to the modification of another one by the second one:



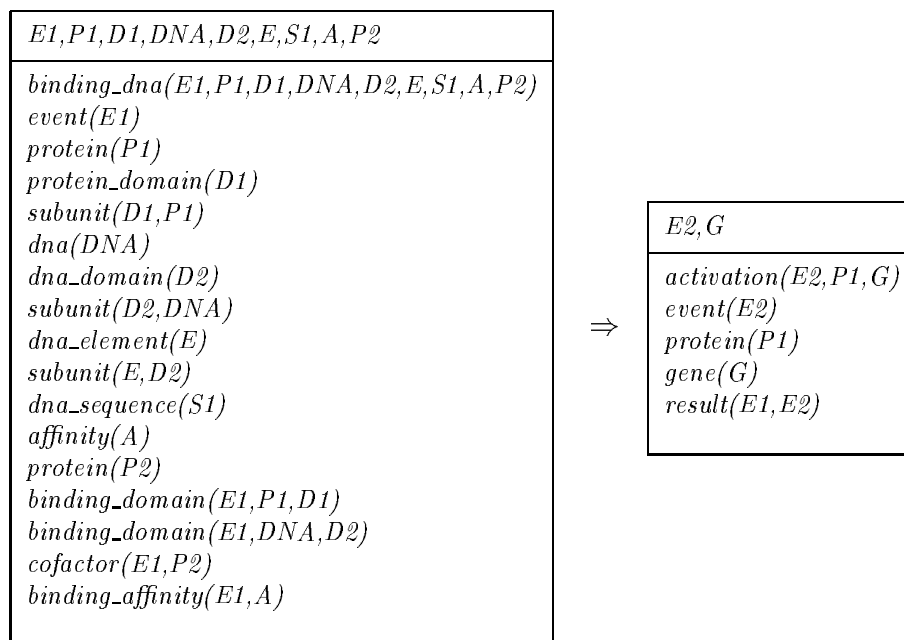
Binding to a gene causes activation, repression or inhibition of it:



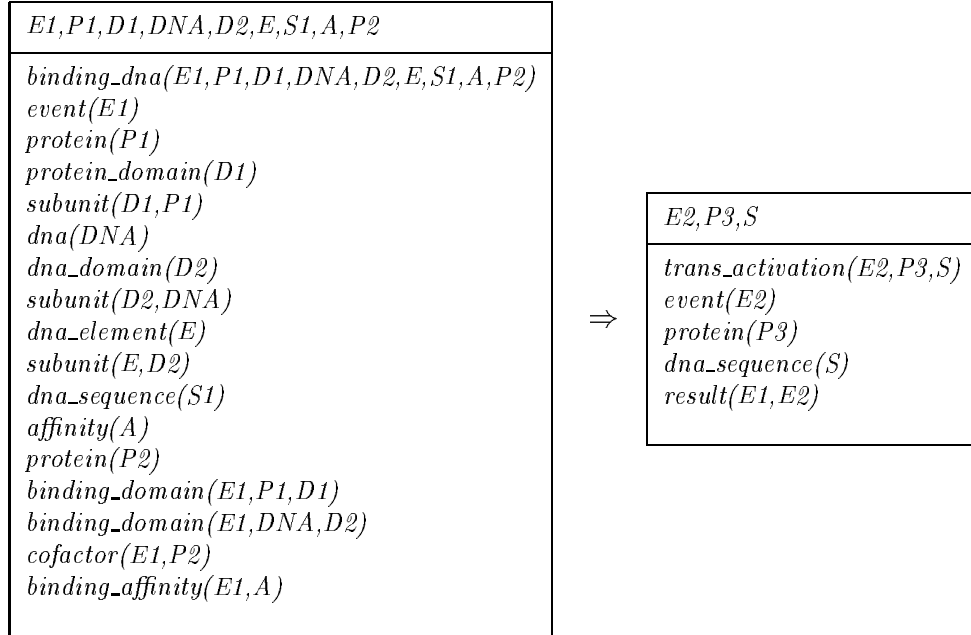
An activation of a gene presupposes a DNA binding:



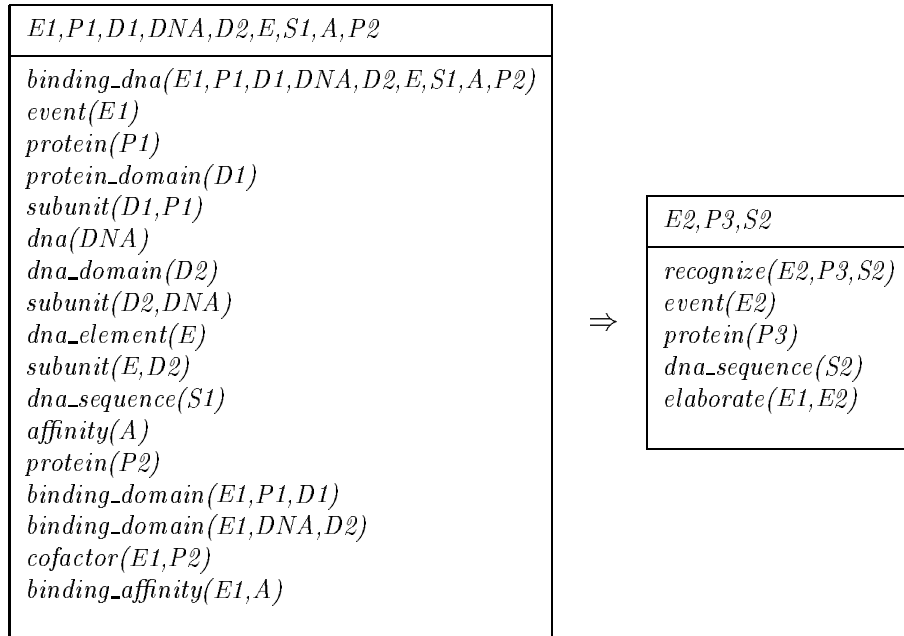
Binding of a protein to DNA causes activation of some gene:



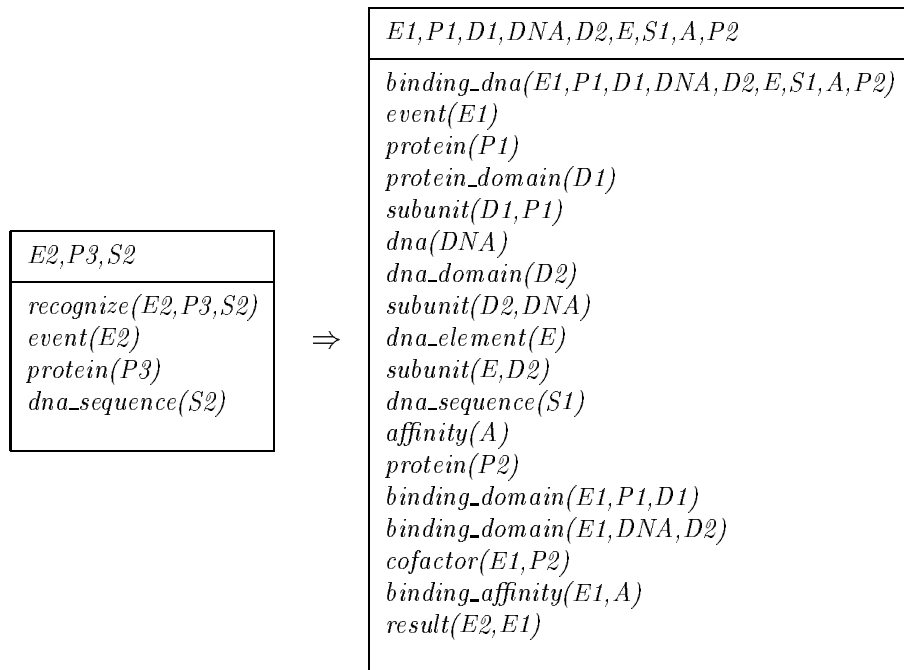
Binding of a protein to DNA causes trans_activation of some sequence:



A protein which binds to DNA recognizes a specific sequence:



The recognition of a certain sequence leads to a binding:



References

- [1] N. Asher and A. Lascarides. Lexical disambiguation in a discourse context. *Journal of Semantics*, 12:61–108, 1995.
- [2] N. Asher and A. Lascarides. Bridging. *Journal of Semantics*, 15:83–113, 1999.
- [3] A. Bairoch and R. Alpweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucleic Acid Research*, (28):45–48, 2000.
- [4] P. Blackburn and J. Bos. Representation and inference for natural language, a first course in computational semantics volume ii, working with discourse representation structures, 1999.
- [5] P. Blackburn, J. Bos, M. Kohlhase, and H. de Neville. Inference and computational semantics. In H. Bunt and E. Thijsse, editors, *Proceedings of the IWCS-3*, pages 5–21, 1999.
- [6] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, 1999.
- [7] J. Bos, P. Buitelaar, and M. Mineur. Bridging as coercive accomodation. In E. Klein, S. Manandhar, W. Nutt, and J. Siekmann, editors, *Working Notes of the Edinburgh Conference on Computational Logic and Natural Language Processing (CLNLP-95)*, 1995.
- [8] T.A. Brown. *Moderne Genetik*. Spektrum, 1999.
- [9] J. Carletta. Aseessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.
- [10] H. Clark. Bridging. In P.N. Johnson-Laird and P.C. Wason, editors, *Thinking, Readings in Cognitive Science*, pages 411–420. Cambridge University Press, 1977.
- [11] H. H. Clark. *Arenas of Language Use*, chapter Common Ground and Language Use, Definite Reference and Mutual Knowledge. CSLI, 1992.
- [12] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
- [13] The Gene Ontology Consortium. Creating the Gene Ontology resource: Design and implementation. *Genome Research*, (11):1425–1433, 2001.
- [14] D. Davidson. The logical form of action sentences. In N. Rescher, editor, *The Logical Form of Action and Preference*. 1967.

- [15] K. Fukuda and T. Takagi. Knowledge representation of signal transduction pathways. *Bioinformatics*, 17(9):829–837, 2001.
- [16] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Towards information extraction: Identifying protein names from biological papers. In *Proceedings of the 3rd Pacific Symposium on Biocomputing*, pages 707–718, 1998.
- [17] R. Gaizauskas and K. Humphreys. Quantitative evaluation of coreference algorithms in an information extraction system. In S. Botley and T. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*, 1996.
- [18] R. Grishman. The nyu system for MUC-6 or where’s the syntax? In *Proceedings of the MUC-6 Workshop*.
- [19] T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In *Formal Analysis in Conceptual Analysis and Knowledge Representation*. Kluwer, 1993.
- [20] U. Hahn, M. Strube, and K. Markert. Bridging textual ellipsis. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING’96)*, pages 496–501, 1996.
- [21] J.A. Hawkins. *Definiteness and Indefiniteness*. Croom Helm, London, 1978.
- [22] J. Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation as abduction. *Artificial Intelligence*, 63:69–142, 1993.
- [23] K. Humphreys, R. Gaizauskas, and S. Azzam. Event coreference for information extraction. In *Proceedings of the ACL/EACL’97 Workshop, Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, 1997.
- [24] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, and B. Mitchell. University of sheffield: Description of the lasie-ii system as used for MUC-7. In *Proceedings the MUC-7 Workshop*.
- [25] H. Kamp. The importance of presupposition. In Christian Rohrer, Antje Roßdeutscher, and Hans Kamp, editors, *Linguistic Form and its Computation*, chapter 7, pages 207–254. CLSI, 2001.
- [26] H. Kamp and U. Reyle. *From Discourse to Logic*. Kluwer, 1993.
- [27] H. Kamp and U. Reyle. A calculus for first order discourse representation structures. *Journal of Logic, Language, and Information*, 5:297–348, 1996.
- [28] H. Kamp and A. Roßdeutscher. DRS-construction and lexically driven inference. *Theoretical Linguistics*, 20(2/3):166–235, 1994.
- [29] P. D. Karp. An ontology for biological function based on molecular interactions. *Bioinformatics Ontology*, 16(3):269–285, 2000.

- [30] Y. Kawazoe and T. Naka. Signal transduction pathways in development: the JAK/STAT pathway. In *Encyclopedia of Life Sciences*. 2001.
- [31] R. Kowalski and M. Sergot. A logic-based calculus of events. *New Generation Computing*, 4:67–95, 1986.
- [32] E. Krahmer and K. van Deemter. Presuppositions as anaphors: Towards a full understanding of partial matches. In P. Dekker, J. van der Does, and H. de Hoop, editors, *De Dag, Proceedings of the Workshop on Definites*, 1997.
- [33] J.R. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.
- [34] A. Lascarides and N. Asher. Discourse relations and defeasible knowledge. In *Meeting of the Association for Computational Linguistics*, pages 55–62, 1991.
- [35] A. Lascarides, N. Asher, and J. Oberlander. Inferring discourse relations in context. In H. S. Thompson, editor, *Proceedings of the 30th Annual Meeting of the ACL*, pages 1–8. Morgan Kaufmann, 1992.
- [36] P. Cimiano Lavin. The role of RNA polymerase in gene transcription from a linguistic point of view. Ms. University of Stuttgart.
- [37] W. Lehnert, C. Cardie, D. Fisher, J. McCarthy, E. Riloff, and S. Soderland. Evaluating an information extraction system. *Journal of Integrated Computer-Aided Engineering*, 1(6), 1994.
- [38] D. Lewis. *Counterfactuals*. Blackwell and Harvard University Press, 1973.
- [39] D. Lewis. Scorekeeping in a language game. *Journal of Philosophical Logic*, (8):339–359, 1979.
- [40] A. Maedche, G. Neumann, and S. Staab. Bootstrapping an ontology-based information extraction system. In J. Kacprzyk, J. Segovia, P.S. Szczepaniak, and L.A. Zadeh, editors, *Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web*. Springer, 2002.
- [41] J.F. McCarthy and W.G. Lehnert. Using decision trees for coreference resolution. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, 1995.
- [42] C. Müller and M. Strube. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 90–95, 2001.
- [43] R. Muñoz and A. Fernández. Resolving definite descriptions in spanish. In *Proceedings of the International Conference on Artificial and Computational Intelligence For Decision, Control and Automation in Engineering and Industrial Applications, ACIDCA2000*, pages 140–145, 2000.

- [44] R. Muñoz and M. Palomar. Semantic-driven algorithm for definite description resolution. In *Proceedings of Recent Advances in Natural Language Processing, RANLP2001*, pages 180–186, 2001.
- [45] R. Muñoz, M. Saiz-Noeda, A. Suárez, and M. Palomar. Semantic approach to bridging reference resolution. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the New Millennium, MT2000*, 2000.
- [46] T. Ohta, Y. Tateisi, and J. Kim. The genia corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference*, 2002.
- [47] N. W. Paton, S. A. Khan, A. Hayes, F. Moussouni A. Brass, K. Eilbeck, C. A. Goble, S. J. Hubbard, and S. G. Oliver. Conceptual modeling of genomic information. *Bioinformatics*, 16(6):548–557, 2000.
- [48] P. Piwek and E. Krahmer. Presuppositions in context: constructing bridges. In P. Brezillon and M. Cavalcanti, editors, *Formal and Linguistic Aspects of Context*. Kluwer, 1997.
- [49] M. Poesio. Scaling up anaphora interpretation. In R. Malaka, R. Porzel, and M. Strube, editors, *Proceedings of the First International Workshop On Scalable Natural Language Understanding*, 2002.
- [50] M. Poesio and U. Reyle. Underspecification in anaphoric reference. In *Proceedings of the IWCS-4*, 2001.
- [51] M. Poesio and R. Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1998.
- [52] M. Poesio, R. Vieira, and S. Teufel. Resolving bridging references in unrestricted text. In *Proceedings of the Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, 1997.
- [53] E. Prince. The ZPG letter: subjects, definiteness, and information-status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. Philadelphia/Amsterdam: John Benjamins B.V., 1992.
- [54] J. Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4):209–441, 1991.
- [55] J. Pustejovsky, J. Castan, and J. Zhang. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Proceedings of the Pacific Symposium on Biocomputing 2002*, 2002.
- [56] U. Reyle. An ontology of events, states and time. Tutorial at the European Media Laboratory.

- [57] U. Reyle. On reasoning with ambiguities. In *Proceedings of the 6th meeting of the EACL*, pages 1–8, 1995.
- [58] U. Reyle. Towards an ontology of biochemical compounds, reactions and pathways. 2002.
- [59] U. Reyle and I. Rojas. A two-dimensional architecture for robust NLP. Technical report, University of Stuttgart, 2002.
- [60] U. Reyle and A. Roßdeutscher. Temporal underspecification in discourse. In C. Rohrer, A. Roßdeutscher, and H. Kamp, editors, *Linguistic Form and its Computation*, chapter 8, pages 255–283. CLSI, 2001.
- [61] U. Reyle and J. Saric. Ontology driven information extraction. In *Proceedings of the nineteenth Twente Workshop on Language Technology*. University of Twente, 2001.
- [62] T. C. Rindflesch, J. V. Rajan, and L. Hunter. Extracting molecular binding relationships from biomedical text. In *Proceedings of the ANLP-NAACL 2000*, pages 188–195. ACL 2000, 2000.
- [63] A. Rzhetsky, T. Koike, S. Kalachikov, S. M. Gomez, M. Krauthhammer, S. H. Kaplan, P. Kra, J. J. Russo, and C. Friedman. A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics Ontology*, 16(12):1120–1128, 2000.
- [64] S. Soderland and W. Lehnert. Corpus-driven knowledge acquisition for discourse analysis. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 827–832, 1994.
- [65] S. Soderland and W. Lehnert. Learning domain-specific discourse rules for information extraction. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.
- [66] S. Staab and A. Maedche. Ontology engineering beyond the modeling of concepts and relations. In R. V. Benjamins, A. Gomez-Perez, N. Guarino, and M. Uschold, editors, *Proceedings of the 14th Conference on Artificial Intelligence, Workshop on Applications and Problem-Solving Methods*, 2000.
- [67] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll. Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on Biocomputing 2000 (PSB'2000)*, pages 541–551, 2000.
- [68] A.M. Turing. Computing machinery and intelligence. *Mind*, 49(236), 1950.
- [69] R. A. van der Sandt. Presupposition: Projection as anaphora resolution. *Journal of Semantics*, (9):333–377, 1992.

- [70] R. Vieira and M. Poesio. Processing definite descriptions in corpora. In S. Botley and M. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*. UCL Press, 1997.
- [71] R. Vieira and M. Poesio. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593, 2000.
- [72] R. Vieira and S. Teufel. Towards resolution of bridging descriptions. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the ACL-EACL'97 Joint Conference: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 522–524, 1997.
- [73] H. Xu. English-style and chinese-style topic: a uniform semantic analysis. In *Proceedings of PACLIC-13, 13th Pacific-Asian Conference on Language, Information and Computation*, 1999.
- [74] A. Yakushiji, Y. Tateisi, and Y. Miyao. Event extraction from biomedical papers using a full parser. In *Proceedings of the Pacific Symposium on Biocomputing*, 2001.
- [75] R. Yangarber and R. Grishman. Nyu: Description of the proteus/pet system as used for MUC-7 st. In *Proceedings of the MUC-7 Workshop*.