Semantic Relations Across Syntactic Levels

by

Viviana A. Nastase

Thesis submitted to the Faculty of Graduate and Post-Doctoral Studies
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

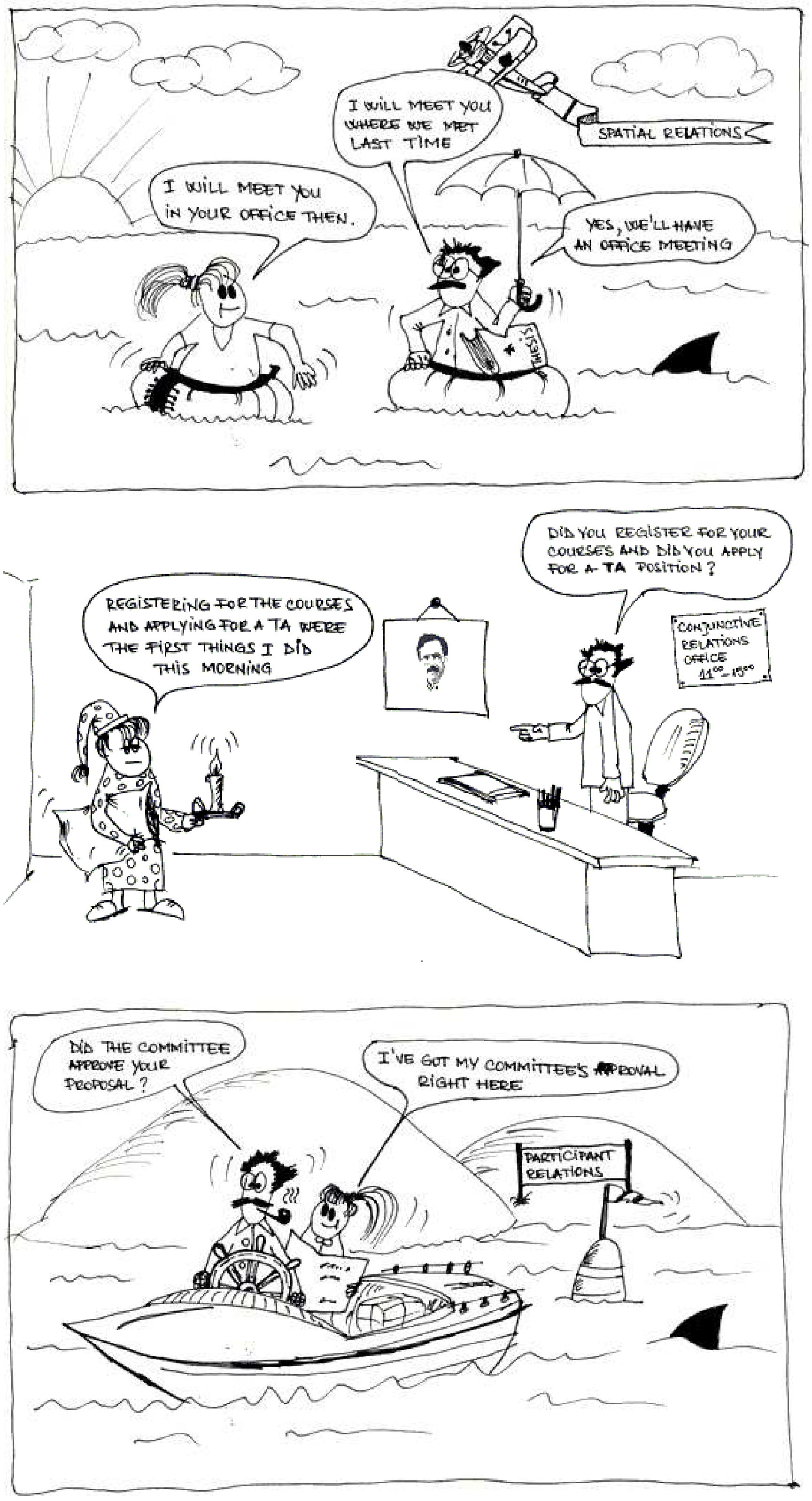
Ottawa-Carleton Institute for Computer Science
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada

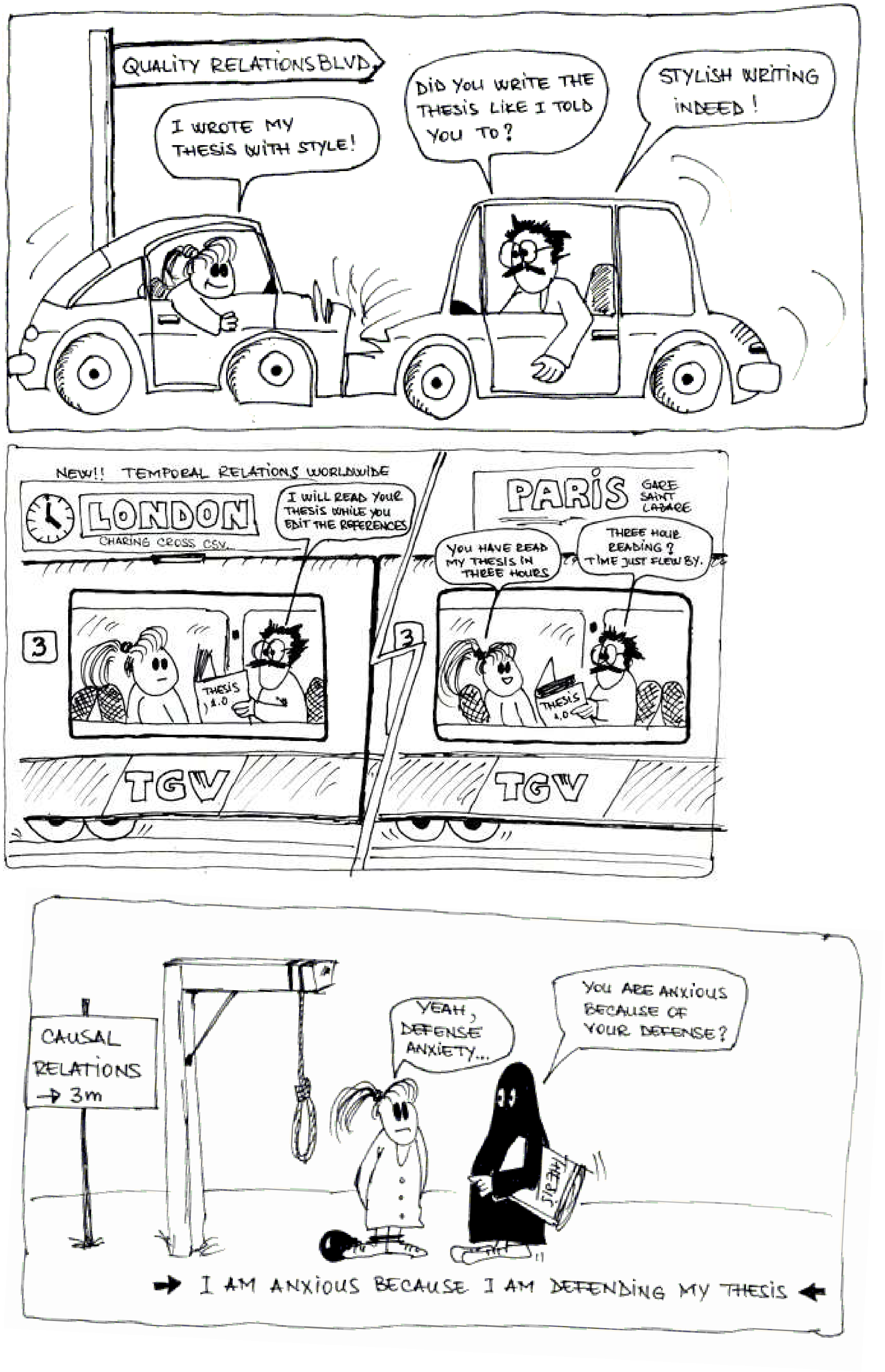
*La connexion est **indispensable** à l'expression de la pensée.*
Sans la connexion, nous ne saurions exprimer aucune pensée continue
et nous ne pourrions qu'énoncer une succesion d'images
et d'idées isolées les une des autres et sans lien entre elles.[1]
Tesnière (1959)

[1] The connection is **indispensable** to the expression of thought. Without the connection, we would not be able to express any continuous thought and we could only list a succession of images and ideas isolated from each other and without any link between them.

To my family

Abstract


SPATIAL RELATIONS
I WILL MEET YOU WHERE WE MET LAST TIME
I WILL MEET YOU IN YOUR OFFICE THEN.
YES, WE'LL HAVE AN OFFICE MEETING
THESIS
DID YOU REGISTER FOR YOUR COURSES AND DID YOU APPLY FOR A TA POSITION ?
REGISTERING FOR THE COURSES AND APPLYING FOR A TA WERE THE FIRST THINGS I DID THIS MORNING
CONJUNCTIVE RELATIONS OFFICE 11^{00}–15^{00}
DID THE COMMITTEE APPROVE YOUR PROPOSAL ?
I'VE GOT MY COMMITTEE'S APPROVAL RIGHT HERE
PARTICIPANT RELATIONS


QUALITY RELATIONS BLVD.
I WROTE MY THESIS WITH STYLE!
DID YOU WRITE THE THESIS LIKE I TOLD YOU TO?
STYLISH WRITING INDEED!
NEW!! TEMPORAL RELATIONS WORLDWIDE
LONDON
CHARING CROSS CSV...
I WILL READ YOUR THESIS WHILE YOU EDIT THE REFERENCES
PARIS GARE SAINT LAZARE
YOU HAVE READ MY THESIS IN THREE HOURS
THREE HOUR READING? TIME JUST FLEW BY.
3
THESIS 1.0
3
THESIS 1.0
TGV
TGV
CAUSAL RELATIONS → 3m
YEAH, DEFENSE ANXIETY...
YOU ARE ANXIOUS BECAUSE OF YOUR DEFENSE?
THESIS
→ I AM ANXIOUS BECAUSE I AM DEFENDING MY THESIS ←

In order to make sense of a message conveyed to us via a spoken or written utterance, we understand what things are talked about, and how they are connected.

From this point of view, do these sentences convey different messages?

I will arrive at 11 am. and *I will arrive when you arrive.*

I will meet you in the office. and *I will meet you where we met last time*

Sweets before dinner spoil your appetite. and *Eating sweets before dinner spoils your appetite.*

I will arrive at a certain point in time: at *11 am.*, or when *you arrive. I will meet you* at a certain place: in *the office*, or where *we met last time.* We can talk about *sweets* and mean *eating sweets.*

Literature review suggests that the relations exemplified by these pairs of sentences are different, because they connect different types of syntactic units. The first relation in each pair connects a verb and one of its arguments, the second – two clauses.

Such distinctions are artificial. Semantic relations link concepts, and will surface on the syntactic level on which the concepts they connect surface.

We aim to give an account of semantic relations that does not depend on syntactic levels. We will justify a unified view of semantic relations across syntactic levels. Such a view has a positive effect on text analysis. It will allow us to gather evidence for a particular semantic relation from all levels at which it appears. Having such information that is not separated according to syntactic levels will allow a text analysis and knowledge acquisition system to use at each processing step, all the evidence previously gathered. We will show that this translates into faster learning and better results.

We can take semantic relation analysis onto another level. We can look for descriptions of concepts connected by a specific semantic relation to find what characteristics or features of the concepts connected make them interact in this way.

blue book, *happy person*, *interesting study*

paper bag, *wooden chair*, *iron gate*

oak tree, *cumulus cloud*, *flounder fish*

Blue, happy, interesting are properties, and *paper, wood, iron* are materials. *Oak* is a specific type of *tree*, *cumulus* is a type of *cloud*, and *flounder* is a type of *fish.*

We will use ontologies to find similarities between concepts that explain or give us indications about the semantic relations in which they are involved.

All these aspects we explore serve to improve text analysis. We propose a uniform processing of texts that allows us to extracts pairs of concepts that interact, and to describe this interaction through semantic relations.

Acknowledgements

I would like to start by expressing sincere thanks to my supervisor, prof. Stan Szpakowicz. He took a chance by giving me this project, and I thank him for his trust. I also thank him for his guidance throughout my years in the PhD program, from close up and from far away. I thank him for his humour and impressive knowledge of matters ranging from NLP to politics, to history, to cookies and beyond, and for sharing these with me.

I wish to thank the members of my committee: prof. Fred Popowich, prof. Jean-Pierre Corriveau, prof. Caroline Barrière and prof. Nathalie Japkowicz for their helpful comments and advice.

I will always remember fondly prof. Ivan Rival, and his wonderful course on ordered sets.

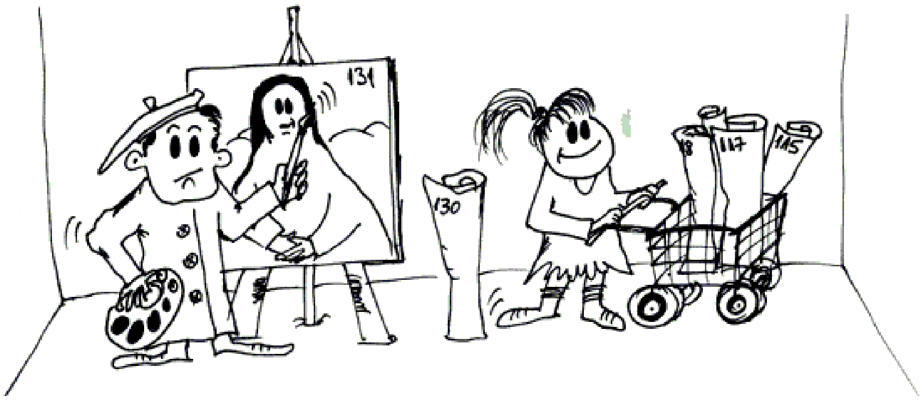
My friends have made my home seem closer, and the time away from my family more fun than I could have hoped for: Igor and Nicholas, the other two musketeers, and my first two friends, always around providing humorous and also serious support and company, Rossana, my dear friend and mentor through my comprehensive exam and also guide through various matters particular to international students, Johanne, my first office mate who showed me around and reminded me weekly our TAMALE seminars, Rimon and Jefferson, my very quiet office mates, Ken and Terry who have answered kindly many many questions, Khalid who introduced me to Perl and shell scripts, Alan, Stephane, Masa, Francisco and Gina, Felipe, my pool playing partners, Nazih with whom I had long and wonderful talks about music and ordered sets, Marina and Mario with whom I exchanged ideas and books many many times, Marina also helped with the examples in Russian in this thesis, Sanda and Rami who brought with them a flavour of my hometown, Fernanda who was very supportive, and many more.

There are a few people whom I need to express special thanks to.

Rada Mihalcea. It is because of her that I am here. I thank her for the fateful message that brought me to Ottawa, for involving me in various projects, and for proofreading my thesis.

Gabi Chindriş. I thank him for his humour and they way it surfaces in images. He managed to make even my serious ideas take a humorous form. Knowing all this, he actually drew his own tribute.

Elisa Paoletti. I thank her deeply for her patience and kindness. She has offered her time and knowledge in language matters in support of my experiments.



Marta Camacho-Zamorra and the whole Zamorra gang. I cannot thank her enough for lifting up my spirit so many times. She is my local sister, and a wonderful, patient and wise friend.

Daniela Savin, Domi and their entire family. My parents join me in thanking Daniela and Domi for caring for me as if I was family. They were always there for me, with support, kind words, countless presents and absolutely delicious food.

Jelber Sayyad. He knows when to be there, and when to be away. He knows when to talk, when to listen, and when to just be there in silence. And he really knows his Perl. I couldn't have had a better friend if I were to design one myself.

And last, as usual, but definitely not least, my family. My Mom, Dad, Cici, Alex and Aunt Florica have supported me in more ways than they realize. They have cheered me up inadvertently so many times. They have cared for me as efficiently from far away as they would have done from up close, and most of all, they have trusted me all along. For all the love they have given me, this thesis is dedicated to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

info
CONCLU
SURFACE FORMS OF MENTAL CONCEPT
SEMANTIC RELATIONS 5km
SEMANTIC RELATIONS ACROSS SYNTACTIC LEVELS

1.1 The Problem

When we want to understand a message conveyed to us through a spoken or written utterance, it is important to understand the things or events talked or written about, and the connections between them. If we consider the words: *John, Mary, loves* separately from the sentence:

1 *Mary loves John.*

we cannot fully understand the message that the sentence conveys. We need to make connections between these elements, and realize that John is the Object of Mary's affection.

Our focus in the research reported in this dissertation is on understanding better these connections, in order to use them in a systematic way to the analysis of texts.

A comprehensive literature review has revealed to us two trends in the analysis of semantic relations.

1. Relations are assigned intuitively to pairs of syntactic units extracted from texts. Analyzing semantic relations from the point of view of the syntactic units they connect anchors them strongly into syntax, and artificially splits them into groups according to the syntactic level to which the units they connect pertain.

2. In the context of creating knowledge representations for information extracted from texts, semantic relations are assigned to pairs of lexical concepts. Semantic relations are a tool that serve a representational purpose.

We will show in Chapter 2 how these two trends in research involving semantic relations arise from the literature.

In order to have a systematic view of semantic relations, we need to bring these two approaches together. In order to do that, we need first to show that semantic relations are the same across syntactic levels. And then we analyze the concepts they connect to find what they have in common.

By comparing various lists of semantic relations in the literature we have observed that lists designed to describe connections between syntactic units from different levels have something in common. There are causal, temporal or other types of relations at different levels. Nobody seems to have explained systematically these numerous commonalities. We are going to investigate this situation, and propose and justify an explanation that semantic relations are in fact the same across syntactic levels.

We approach this problem by explicitly taking the position that **semantic relations** describe interactions between concepts, and not syntactic units. We then look for a bridge between concepts and syntactic units.

When we use the term **concept**, as it will be further discussed in Chapter 3, we refer to the things we talk about: occurrences, entities, and attributes of occurrences and entities. Concepts underlie not only words, but also larger syntactic units. The definition that we adopt shows a coarse-grained view of concepts, which we will compare in Section 3.2 with more fine-grained views. We choose such a coarse view because it allows us to find mappings between concepts in our mind, and the syntactic structure of the expressions that evoke these concepts. Finding pairings between concepts and syntactic structures will allow us to show that one concept can take different forms, for example, an occurrence can be expressed through a clause or possibly just a noun. We consider structures pertaining to three syntactic levels: clause, intra-clause and noun-phrase.

Concepts interact in few ways, while the number of pairs of concepts that interact is practically unlimited. As an example for this idea, let us consider the Material relation. This relation describes the connection between an artifact and the material it is made of. There are countless examples of such a relation: *wooden door, iron fence, plastic bottle*, and so on. Undoubtedly, pairs of concepts connected by the same relation must have something in common. There are two ways in which we can analyze what the corresponding concepts share:

1. Find whether they have a common ancestor in an ontology (which means that they share some common traits).

2. Reverse the process and start from analyzing semantic relations and the constraints they impose on the concepts they connect. These constraints will be shared by all concepts connected by a specific relation.

We explore the first approach in Chapter 6. This is the approach adopted by researchers concerned with knowledge representations, like conceptual graphs. Concepts are represented as collections of primitives, and the semantic relations are defined as linking concepts that share certain primitives. In order to have a computationally viable system that assigns semantic relations based on conceptual primitives, one would need a resource that describes all concepts in terms of primitives. Such a resource is very labour-intensive, and it encounters the design problem of establishing a set of necessary and sufficient primitives to be used for the description of all concepts. Theories of concepts based on conceptual primitives do not yet agree on such a set of features, and the properties chosen to describe concepts that seem adequate from one perspective, may be inadequate from others (for example concepts in euclidian relative to alternative geometries) (Lormand, 1996), (Giaquinto, 1996), (Kripke, 1980), (Putnam, 1975). Instead of using hand-written definitions of the type of concepts that semantic relations connect, we use ontologies to find what types of concepts each semantic relation that we analyze connects. The methodology can be applied to any semantic relation.

For the most part of this dissertation we focus on the second approach, which will lead us to a unified view of semantic relations across syntactic levels.

From a superficial point of view, semantic relations do link syntactic units. The utterances that we use to convey a message consist of words which are not randomly put together, but follow certain constraints. These constraints are captured by grammatical rules. We can superimpose grammatical structure onto any (correct) linguistic expression. The relations should not depend on the particular structure that an expression has, but on the concept conveyed by this expression. We analyze relations from the point of view of the concepts they connect, and see how these concepts can be mapped onto different structures.

Let us take an example. TimeAt is a relation that connects an event/state/activity/etc. with a point on the time axis, which indicates when the event/state/activity/etc. has taken place. So one of the concepts linked must be an occurrence, while the other must indicate a time point. How can we express a time point? One way is to be explicit, and say for example *11 am.*, or *11 am. on October 17th 2003.* There are other ways to indicate time. Occurrences have an implicit time frame – occur, start unfolding, finish unfolding at a certain time or interval. We can use the time frame of an occurrence as a temporal reference for another. In order to express a time point, we choose a punctual event, for example *he finished reading the article.* The two sentences:

2 *He left at 11am.*

3 *He left when he finished reading the article.*

both show instances of the TimeAt relation, between the concept of he leaving and the time point 11am, and between the concept of he leaving and the time point he finished reading the article.

Also, an occurrence may have different expressions. The clause *he left* represents a leaving occurrence, just as the gerund *leaving* does. Then the sentences in examples (4) and (5) will display the same TimeAt relation as sentences (2) and (3).

4 *leaving at 11am*

5 *leaving when he finished the article*

We observe that a concept can take various syntactic forms, in other words, can surface at different syntactic levels. Semantic relations connect concepts. A relation will have instances on the syntactic levels on which the concepts they connect have surfaced.

When we talk about different surface forms of concepts, we recognize that there are differences that we do not analyze. We will say that the same concept[1] underlies different expressions, from the perspective that it evokes the same general idea, or conveys the same message, in the mind of the listener/reader. Also, when we analyze mappings of concepts onto surface forms, we always perform such mapping in the context of a pair concept and the semantic relation that links them. The surface forms by themselves, without the context, would not convey the same message.

1.2 The Working Assumptions

We approach semantic relation analysis from the point of view of computational text analysis.

We place the following restriction on the text analysis that we perform: to require as little manual labour as possible in terms of lexical resources. This restriction will actually give more liberty in terms of the texts we can process. We do not restrict the vocabulary to a certain domain, and we are not making text analysis depend on semantic resources that are labour-intensive, domain-dependent and hard to obtain. We build our own resources while processing texts, by accumulating knowledge from the input. At this point, all we use is a dictionary of English words that gives part-of-speech information, a list of function words (prepositions and connectives) that describe the semantic relations that these words may indicate and a syntactic parser that builds parse trees for input sentences.

In order to be able to focus on semantic relations between concepts expressed in texts, we make some simplifying assumptions about the type of input data that we analyze.

We focus on semi-technical texts – expository texts which contain little or no figurative speech, and which can be understood in their literal sense (Copeck et al., 1997). Since a semi-technical text is considered to be an objective account, the truth value of the sentences it contains is not in question.

Metaphors are figures of speech that allow us to express novel ideas through analogies with established ideas (Lakoff, 1987). There are ongoing attempts at building common-sense knowledge bases (Lenat, 1995),(Singh *et al.* (2002)) that a computer could use to make the type of inferences and analogies that people make. The effort of creating these resources was bigger than the initiators had anticipated, as they are not yet large enough

[1] In a more precise definition of meaning, we could only talk of a similar, or very similar, concept.

to be used in a system that processes general texts, with unrestricted vocabulary and little hard-coded semantic information, as is our aim.

The truth value of an utterance is also something that can be questioned from a philosophical point of view. Sentences are decomposed into propositions whose truth value is used to find the value of the sentence that encompasses them. Propositions may be **necessarily** true – which means that they are true in all possible worlds, as the German philosopher Gottfried Leibnitz would put it (Lyons, 1995, p. 118) – or **contingently** true – the proposition may have different truth values in different worlds. Necessary truths can be **analytical** – analysis of the arguments of a proposition can reveal its truth value – or **logical** – in which establishing the truth value is based on the logical system used. Logical truths are considered to be a subset of analytical truths. The problem can be further complicated by distinguishing **epistemologically true** propositions – the ones that are true given on our present knowledge, but may be proven false by further analysis – and propositions whose truth value was established by analysis (Lyons, 1995).Quantifiers, mode and tense play a very important part in sentence analysis from a truth-value point of view.

All these aspects would interfere with our purpose of analyzing texts from the perspective of the concepts evoked and the way they interact. We will focus on extracting units from texts that are surface expressions of concepts, and then link such units by means of semantic relations that describe the way they interact.

Semantic analysis, and semantic relation analysis in particular, is a more subjective matter than part-of-speech, morphologic and surface-syntactic analysis. We will try to anchor the subjective part of our text analysis into the more objective aspects, in particular syntax. Native speakers of a language are rarely aware of structural rules and restrictions they apply when producing language. It is accepted in cognitive circles that since there is an infinite number of correct sentences that can be produced, and a finite number of lexemes, there must be some underlying system, which may function subconsciously for native speakers of a language that allows us to form sentences (Pinker, 1999), (Pinker, 1995). This system would capture the structure in language, and it is the focus of grammatical studies. Theories of grammar, and of the universality of grammar, have been around for a very long time. Panini's work on Sanskrit, *Astadhyayi*, from around the 5th century BC, shows the oldest documented comprehensive account of morphology, phonology and grammar. Together with other ten grammarians (as mentioned in his work) he proposed 4000 rules, expressed as aphorisms that describe the structure of the Sanskrit language (Joseph, 1991),(Cardona, 1976).

It may be debatable whether syntax is a valid starting point for text analysis. Attempts were made to analyze text without using syntax. Schank (1975) for example proposed a model based on cognitive research. This research shows that we process the sentence word by word, and each new word builds expectations in the recipient of the message, which should be fulfilled by the subsequent words. The shortcoming of such an approach, from a computational point of view, is that it requires a large manually build resource that describes each word in terms of the expectations it generates.

It can also be argued that syntax is important for languages with relatively fixed word order, like English. Since the role of a word in the sentence is rarely marked by case endings, as it often happens in free-order languages, we rely on syntax to reach the meaning of a message. For example, the sentences:

6 *John loves Mary.*

7 *Mary loves John.*

use the same words, but convey completely different messages. We can account for the difference in meaning through the difference in syntactic structure: in the two sentences, the words play different grammatical roles.

Syntax, from the point of view of semantic analysis research, has other advantages as well. It provides us with structure and with indicators – prepositions, connectives – that allow us to find syntactic units that interact, and give clues about the way in which such units interact.

The fact that syntax helps semantic analysis is further supported by NLP work that proposes a syntax-first approach to text analysis. Some of this work is concerned with knowledge acquisition (Ahlswede and Palmer, 1988),(Klavans et al., 1992),(Briscoe, 1991), (Copeck et al., 1992) or information extraction (Shinyama et al., 2002).

Because in the text processing that we perform we rely on syntax, we have to place another restriction on this analysis. Although texts are structured into chapters, sections, paragraphs and sentences, the levels above the sentence have not been formalized enough to give us the same indicators and structural information that the sentence level does.

Below sentences and phrases there is the level of morphology. We only look at it briefly, in view of our work on semantic relations inside a word (Nastase and Szpakowicz, 2003a). We leave a more thorough analysis for future work.

We focus on processing sentences and units inside a sentence. Throughout our analysis we do not commit to one particular theory of grammar, to maintain our discussion on a more general level. We distinguish the structural units we use based on the open-class word that dominates the structure: verb-centered (correspond to clauses), nominal-centered (noun-phrases), adjective- and adverb-centered.

As mentioned before, the focus of this dissertation are semantic relations, and using them to improve text analysis. We started by studying three lists of semantic relations, one for each of the following syntactic levels: clause, intra-clause and noun-phrase. They were designed to be used in a semi-automatic knowledge acquisition system, to comprehensively process semi-technical texts. Each list was designed for a specific syntactic level, thereby linking different types of syntactic expressions. A superficial analysis has shown some overlap in these lists, and other lists discussed in the literature. The reason why we observe the overlap may be that it has been assumed that semantic relations are the same, and that a causal relation for example can be assigned to a pair of clauses that interact, or a noun and its modifier. Another reason may be that in naming an interaction between syntactic units there has been no consideration of deeper implications of assigning the same name to relations that link units pertaining to different syntactic levels. Except for cases (relations between a verb and its arguments) which were transferred to noun-modifier relations because of verb nominalizations, the process of assigning relations with the same name to pairs of syntactic units from different syntactic levels seems ad-hoc. We will make our hypothesis that semantic relations are the same across syntactic levels, and this dissertation will present evidence to support it.

The phenomenon in focus here can be exemplified like this: knowing that a semantic relation describes the interaction between an occurrence and one of its participants, try to account for all possible syntactic manifestations of such a relation. We will see that an occurrence can be expressed either through a verb, a deverbal noun, a deverbal adjective or even a non-deverbal noun. In each such case, the participant in the occurrence will take the syntactic form that allows it to be an element in the structure that expresses the occurrence.

1.3 Concepts and Relations

The meaning of a sentence can be analyzed from a truth-based perspective ((Tarski, 1983),(Tarski, 1944),(Davidson, 1984), etc.), or a conceptual one, by identifying the entities named in the sentence and establishing links between them. Katz and Fodor (1963) set the precedent for this type of semantic analysis, later named compositional, using the notions of deep structure and surface projections from the generative grammars.

According to the compositional paradigm, the meaning of an utterance can be assembled from the meaning of the syntactic constituents of the utterance. The semantic analysis is based on the syntactic structure of the utterance, and it aggregates the meaning captured by smaller units based on this structure (Katz, 2002, in (Margolis and Laurence, 2002)).

We adopt from the compositional paradigm the idea of assembling the meaning of a sentence by analyzing its components. The components that we are interested in are the concepts behind syntactic units, and we put them together by linking them through semantic relations that show how these concepts interact.

We single out this part of the meaning – concepts and the relations between them – but it does not imply that everything else is discarded. All the information regarding the syntactic and morphological form that a concept took is still there, but pushed to the background, to allow us to focus on the semantic relations between concepts.

As we will show in Chapter 2, semantic relation analysis concentrates on the interaction between structures pertaining to specific syntactic levels: two clauses, the main verb and its arguments, or the head noun in a noun phrase and its modifiers. It is probably implied, but not stated explicitly, that semantic relations hold between the concepts expressed through these structures. When we explicitly adopt the view that semantic relations hold between concepts, and that these concepts may take various forms that project them onto specific syntactic levels, we arrive at a unified view of semantic relations across syntactic levels. Semantic relations will not be tied to specific levels. For example, Agent will not be a relation between a verb and one of its arguments. Instead it will be a relation between an occurrence and a participating entity. If the occurrence is expressed through a verb, then the relation will surface as a case, or thematic role. If the occurrence is expressed through a deverbal noun, then the relation will surface as a noun-modifier relation, as shown in examples (8) and (9):

8 (IC) *The parents approved [his choice].*

9 (NP) *parental approval*

Just as a pair of concepts in a given context determine the way they are connected (for example: *wooden lodge* – we perceive a Material connection between the concept

WOODEN (made of wood) and LODGE (building)), the connection itself imposes certain restrictions on the nature of the linked concepts (for example, a CAUSALITY relation requires the concepts connected to be occurrences).

We look closer at what it means for two concepts to interact in the manner described by some relation, and what those concepts should be, to make this description accurate. When people assign a label to such a connection, more knowledge figures in this process than the mere words involved. The words have connotations and bring up associations, few of which are available to computer systems (Lenat, 1995),(Singh et al., 2002). But by recognizing that such information must be there for two concepts to be connected in a certain way, we can prepare a language processing system to look for and find such information as it collects instances of semantic relations from texts.

We will give an example to clarify this idea. CAUSE is a relation that indicates that two concepts, the *Cause* and the *Effect* (which are occurrences), interact in the following manner: the realization of the concept expressing the *Cause* is necessary and sufficient to bring the *Effect* into existence, and the *Cause* is known to exist.

10 *The student was anxious because he was writing an exam.*

This would be an instance of CAUSE at the clause level. The *Cause* occurrence is *write an exam*, the *Effect* is *the student is anxious*. The causal link between the two occurrences has an overt lexical representation: "because". The two concepts in the causal relation are also explicitly expressed as occurrences.

Now, consider the sentence:

11 *The student was anxious because of the exam.*

The relation between the state of *the student being anxious* and the *exam* is still CAUSE. In this case, though, there are not two overt occurrences. The *Effect* is still expressed as an occurrence, while *Cause* is a noun.

We know that CAUSE is a relation that involves two occurrences: the unfolding of one causes the other one to take place (or become true). How can we say that in example (11) we still have the CAUSE relation? When people make this judgment, they bring other information into the sentence, based on their background knowledge, and on trying to make sense of what the speaker wanted to convey through a certain utterance. It is then clear that some action associated with *exam* takes place, and it causes the student to be anxious.

By assigning the appropriate relation to these pairs of concepts, we can try to gather evidence for information that is missing in certain instances. By analyzing a corpus, we might discover several actions associated with *exam* (writing, answering, thinking about), in the context where these actions are connected to a state of *anxiety* by a causal relation. Should appropriate lexical resources, especially ontologies, be available, this could be taken one step further to find generalizations such as:

expository action involving *examination type* Cause *negative emotion*

1.4 Text Analysis and Knowledge Acquisition

Semantic relation analysis can contribute to many NLP tasks, including word sense disambiguation and text summarization. Two major areas of application, however, are information extraction and knowledge acquisition.

Message Understanding Competitions (MUC) (Sundheim, 1993), (Sundheim, 1995) were organized as "competitions" for systems that extract from texts accurate information on events and entities. Such systems would identify a specific event and the entities involved and the roles they play in this event, or alternatively a specific entity, and the events it is involved in and the roles it plays in these events.

Knowledge acquisition is in a way similar to message understanding, but instead of focusing on finding a particular piece of information, it is interested in finding particular types of information or all information from the input available.

Knowledge acquisition can take many forms, and it depends on what the goal of the acquisition process is. There are approaches that build ontologies based on information extracted from dictionaries (Chodorow et al., 1985),(Rigau et al., 1998), or the focus may be on general knowledge extracted from texts (Richardson et al., 1998),(Copeck et al., 1992) or machine-readable dictionaries (Lauer, 1992),(Barrière, 1997),(Ide and Veronis, 1994),(Klavans et al., 1992).

We will work with a semi-automatic text analysis and knowledge acquisition system, TANKA (Copeck et al., 1992), which extracts all pairs of concepts from a general, semi-technical, input text, and assigns each pair a semantic relation that describes how these concepts interact. In the context of this knowledge acquisition system, we will verify that the theoretical principles that explain why semantic relations are the same across syntactic level have positive practical influences. Based on these principles, we will combine three lists of semantic relations that each covers a separate syntactic level, into one list that spans all three levels. We will test if the knowledge acquisition process implemented by TANKA improves when the system uses a combined list and processes texts in a uniform manner without discriminating between syntactic levels, as opposed to the situation when it uses three separate lists, and each syntactic level is addressed separately.

Since semantic relation analysis is useful for knowedge acquisition, we draw from here another support for the idea that syntactic levels should not be a factor in semantic relation analysis. Syntax helps us reach meaning by giving us access to structures that convey concepts and by giving us clues about the way in which the concepts behind these structures interact. Once we get to the concepts and establish how they are connected, syntax can be disregarded. Therefore discriminating semantic relations based on syntactic levels is artificial.

The text analysis and knowledge acquisition system, TANKA (Copeck et al., 1992), has two components:

- a syntactic analyzer – DIPETT – which implements a grammar for English based on the comprehensive theoretical analysis presented in Quirk *et al.* (1985). For each input sentence this parser produces tree structures very rich in syntactic and morphological information (Delisle, 1994).

- a semantic analyzer – HAIKU – which uses the parse trees produced by DIPETT to perform semantic analysis separately on three syntactic levels: clause, intra-clause and noun-phrase (Barker, 1998),(Delisle, 1994).

For each of the syntactic levels analyzed, the system has a separate list of semantic relations. The relations were grouped into three lists, because depending on the syntactic level for which they were designed, they connect different types of syntactic units: clauses in a multi-clause sentence, elements of a clause, and the head of a noun-phrase and its modifiers. Also, each syntactic level provides its own clues for semantic relation analysis – connectives at the clause level, prepositions at the intra-clause and the noun-phrase level.

Tables 1.1 - 1.3 present the lists of relations as developed by Ken Barker and Sylvain Delisle ((Barker, 1998), (Delisle, 1994)).

At a first glance there seem to be some commonalities between these lists – there are relations on each level to express causality, for example. This observation provided the initial motivation for the present research. We have analyzed semantic relations from the point of view of the concepts they connect, to discover whether there are reasons to propose a unified view that does not tie semantic relations to specific syntactic levels.

Table 1.1: The clause level relations

CAUSAL				
Causation	Detraction	Enablement	Entailment	Prevention
TEMPORAL				
Co-occurrence	Precedence			
CONJUNCTIVE				
Conjunction	Disjunction			

Table 1.2: The cases

PARTICIPANT			
Accompaniment	Agent	Beneficiary	Exclusion
Experiencer	Instrument	Object	Recipient
CAUSALITY			
Cause	Effect	Opposition	Purpose
SPACE			
Direction	LocationAt	LocationFrom	LocationThrough
LocationTo	Orientation		
TIME			
Frequency	TimeAt	TimeFrom	TimeThrough
Timeto			
QUALITY			
Content	Manner	Material	Measure
Order			

Table 1.3: Noun-modifier relations

Agent	Beneficiary	Cause	Container
Content	Destination	Equative	Instrumental
Located	Location	Material	Object
Possessor	Product	Property	Purpose
Result	Source	Stative	Time
Topic			

The theoretical part of this dissertation will present an analysis of surface forms of concepts, and phenomena that relate various forms of the same concept. Although we constrain our analysis to the level of the sentence and the structures within it, the analysis we present is general. When there appears a formalized approach that gives us structure and indicators for larger text units just as we have them for sentences, this analysis can be extended to such units. It may be the case that the semantic relations we propose are not adequate to describe interaction between concepts expressed in paragraphs and sections of a text. In that situation, based on the concepts and semantic relations inside these units, it may be possible to find chains of semantic relations, or inferences, that connect the concepts in question. We will propose this possibility for future work.

Because concepts can surface in forms pertaining to different syntactic levels, and semantic relations link concepts, semantic relations will surface on the level onto which the concepts they connect surface. We obtain a view of semantic relation that is independent of syntactic levels. For example, even if a semantic relation will superficially link a noun and its adjectival modifier, on a deeper semantic level it may link two occurrences.

1.5 Goals

The research presented in this dissertation was motivated by the goal of our research group, which is to process and extract knowledge from semi-technical texts in the form of concepts linked by semantic relations. The system which we are gradually building should perform this automatically, while relying on few resources, so that it is not domain dependent. The system should build its own resources as it processes texts.

Our goals arise from this endeavour, and present the theoretical and practical aspects of the same idea: semantic relations are the same across syntactic levels.

The theoretical part will present an analysis of the way in which concepts can surface through syntactic structures inside a sentence. We analyze the relation between surface forms of the same concept, and propose a systematic account of phenomena that explain this relation. We investigate the nature of semantic relations and the type of concepts they connect. Based on the investigation of surface forms of concepts, we show how semantic relations that link occurrences, for example, can have instances on several syntactic levels.

The practical goal is to improve text analysis from the point of view of semantic relations. We propose a new implementation of the semantic analysis module of TANKA which processes texts uniformly, without differentiating between syntactic levels. This type of processing is imposed by the unified view of semantic relations that the theoretical part proposes. To show that this type of processing is beneficial, we compare the performance of TANKA when it uses the two different semantic analysis modules. We measure the learning rate of the system in the two configurations, and observe that in the experiment conducted using a combined list of relations and uniform treatment of texts, the system learns faster – it will make better suggestions earlier in the process.

1.6 Organization of the Dissertation

The dissertation focuses on semantic relations, and the concepts they connect. Chapter 2 justifies our compositional approach to text analysis, and presents the way in which semantic relations have appeared and were incorporated in language analysis. We focus in this chapter on showing that research in semantic analysis does not investigate the type of concepts that semantic relations link. The literature review will show that the pervasive view of semantic relations is strongly anchored in syntax. There are two main currents: focusing on specific syntactic levels and analyzing relations between structures pertaining to that particular level, and focusing on words but not on the type of concepts they represent (occurrences, etc.).

Chapter 3 justifies first of all the definition of concepts we adopt. We present different views of what concepts are considered to be from a cognitive linguistic point of view, and show that our own view does not affect in any way any of these alternative views. We are interested in finding relations between different surface forms of the same concept. This causes us to investigate language production – how concepts are projected from their mental form onto expressions in a language. We look at psycholinguistic research to find clues about the way humans produce language. We look at technical approaches of language generation inspired by psycholinguistic research. Transformational grammars are the ones that explicitly claim to be able to produce different surface forms from the same deep structure. We look at this field to see if there are accounts of phenomena that allow for a concept to take forms pertaining to different syntactic levels. Finally, we turn to the process of word formation, which explains phenomena of paraphrasing and compaction of expressions into single words. Once we have concluded the search for clues about the way in which concepts surface in language in these fields, we turn to texts and analyze different expressions of concepts that are similar in the context of the sentence in which they appear, to find our own account for why concepts can surface in different syntactic forms. We identify and then discuss each phenomenon discovered.

Chapter 4 uses the discussion in Chapter 3 to defend the idea that semantic relations are the same across syntactic levels. Semantic relations hold between concepts, not syntactic units. If the same concept has equivalent expressions (relative to the context in which it appears) pertaining to different syntactic levels that does not affect the relations in which it is involved. We present such considerations for six groups of semantic relations: **causality**, **temporality**, **spatiality**, **conjunctive**, **participant** and **quality**. We focus on these groups because they appear throughout the literature on semantic relation analysis.

In Chapter 5 we test the practical impact of the theoretical ideas presented, in the context of a knowledge acquisition system. Based on the aspects explored in Chapters 3 and 4, we build a list of semantic relations that spans the three syntactic levels we analyze. This list combines three lists that TANKA, the knowledge acquisition system we use, originally used. We show that by using a combined list of relations, the knowledge acquisition process improves. This is revealed by comparing the learning curves that show the behaviour of the system when it uses one list of relations, and when it uses three separate lists.

In Chapter 6 we investigate similarities between concepts connected by the same semantic relation, using ontologies. Our analysis is empirical, and we use machine learning techniques. Similarities between concepts can also help in automating the process of semantic relation assignment.

Up to this point the system has complied with the initial restrictions that our research group has placed on a text analysis and knowledge acquisition system: that it uses few semantic resources, in order to require as little manual labour as possible. The purpose is to allow the system to build its own resources as it processes texts. The project was initiated in the late 1980s. Since then, much has changed in the NLP community. Semantic analysis relies on easily accessible corpora and on public-domain machine-readable lexical resources, notably *WordNet*, as well as on less widely available but extremely promising dictionaries and thesauri, for example *Roget's Thesaurus* or *LDOCE*. The experiments presented in Chapter 6 show that these resources allow us to find rules for semantic relation assignment, and it would be interesting to incorporate them in our text analysis and knowledge acquisition system.

Chapter 7 presents ideas that stem from the current research to be addressed in the future.

Chapter 8 reviews the work done, and extracts conclusions from the theoretical and experimental work presented in this dissertation.

1.7 Notations and Definitions

1.7.1 Definitions

Occurrence encompasses all types of events, actions, activities, processes, states and accomplishments (Allen, 1984). We consider an occurrence to be not only something that is usually expressed through a verb, but the whole situation, including the participants and its attributes. We consider the occurrence to be a *whole*, and its participants and attributes its parts.

Entity covers concrete and abstract objects, everything that usually surfaces in language as nouns.

Attributes describe properties of occurrences and entities. They are usually expressed through adjectives and adverbs.

Concept encompasses occurrences, entities, attributes of occurrences and entities. Concepts are what underlies a linguistic expression, the idea in the mind of the utterer before he projects it into a verbal form, and the idea in the mind of the listener after he has processed this form.

Semantic relation is the link we perceive between concepts in a message that is being conveyed to us, or that we are trying to convey. It describes the way in which we understand the contribution of the concepts toward the meaning of the message.

Semi-technical text is a text which contains little or no figurative speech, and which can be understood in its literal sense (Copeck et al., 1997). The truth value of the sentences in this text is not in question.

1.7.2 Notations

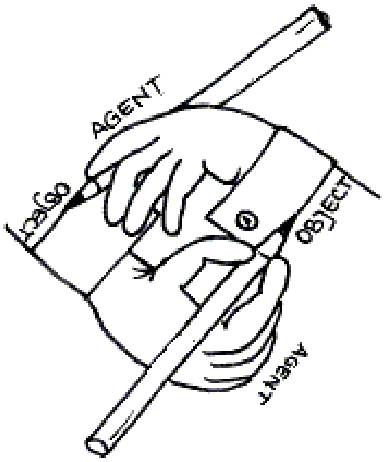
We adopt the following notational convention:

- Concept names will be written in capital letters (e.g. SWEETS).
- Semantic relation names will be written in small caps (e.g. CAUSE, AGENT).
- We distinguish between the semantic relation and the roles implied by the relation (for example, an AGENT relation implies that one of the concepts linked fills an *Agent* role, as is the case for *Mary* in the sentence *Mary is reading*). The roles will be written in italics.

- Classes of semantic relations are written in bold (e.g. **Causality** will encompass all relations that express causal interaction between occurrences).
- CL – Clause Level
- IC – Intra-Clause Level
- NP – Noun Phrase Level
- CLR – Clause Level Relation. We use it to refer to relations between clauses.
- Case – Used to refer to relations inside a clause.
- NMR – Noun-Modifier Relation. Used to refer to relations at the noun phrase level.

Chapter 2

Semantic Relations



2.1 Introduction

This chapter presents a justification for our decision to concentrate on semantic relations between concepts in the semantic analysis of texts.

We start by quoting a few influential personalities in linguistics, who support the view that the meaning of a text comes from the links we establish between the entities mentioned, and not only from the entities themselves.

Once our approach has been justified, we present a historical view of the development of semantic relations. We will see how the semantic relations have appeared and evolved, and how they are used in semantic analysis of texts. We will see that they are strongly linked to syntactic levels. Research concentrates on identifying and analyzing relations between certain types of syntactic units – a verb and its arguments, a noun and its modifiers, two clauses in a multi-clause sentence, sentences or paragraphs in a text.

A unified view would arise naturally, should one consider that semantic relations connect concepts, and not syntactic units. Syntactic units are ways to express concepts, and different structures can convey the same, or a very similar, concept in certain contexts.

2.2 Justifying the Emphasis on Semantic Relations

The importance of relations between words in conveying the meaning of the text, as opposed to analyzing them in isolation, has been recognized by linguists at different points in the development of linguistic analysis.

Whorf (1956, p. 67) expresses such a view:

> It is not words mumbled, but RAPPORT between words, which enables them to work together at all to any semantic result. It is this rapport that constitutes the real essence of thought insofar as it is linguistic.

The RAPPORT between words that Whorf mentions is a term that encompasses many links (syntactic, lexical, semantic).

Tesnière (1959, p. 11-12) expresses a similar view that emphasizes the role of the connection between words, as opposed to the meaning of the words by themselves:

> Tout mot qui fait partie d'une phrase cesse par lui-même d'être isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des **connexions**[1], dont l'ensemble forme la charpente de la phrase. Ces connexions ne sont indiquées par rien. Mais il est indispensable qu'elles soient aperçues par l'esprit, sans quoi la phrase ne serait pas intelligible. [2]

Tesnière views the sentence-meaning compositionally. He defines the subject, object and indirect object non-heterogeneously as "actants" participating in a theatre-like act.

The compositional approach to sentence analysis has received support also from psycholinguistics.

[1] The emphasis is Tesnière's.

[2] Each word which is part of a sentence ceases by itself to be isolated as in a dictionary. Between it and its neighbours, the mind perceives connections, the ensemble of which forms the framework of the sentence. These connections are not indicated by anything. But it is indispensable that they be perceived by the mind, without them the phrase would be unintelligible.

Van Dijk and Kintsch (1983) have discovered that discourse understanding leaves three distinct traces in memory, each corresponding to a different level of processing:

1. **Surface form** – a syntactically, semantically and pragmatically interpreted sequence of words.

2. **Meaning** – an interconnected network of ideas.

3. **The situation referred to** – it is similar to the representation that would result from directly experiencing the situation that the discourse describes.

Experiments have shown that each of these levels has a different representation, and each results from a distinct level of language processing. The three levels have different access times and are differently influenced by sentence boundaries, co-reference, the determinacy of spatial description and other variables. In terms of memory, each has a different decay rate. The surface form has the shortest lasting time, the meaning a longer one, and the situation model the longest.

Van Dijk and Kintsch (ibid.) postulate that the meaning is actually a propositional base. The notion of proposition was borrowed from philosophy and linguistics (Anderson and Bower, 1973), (Kintsch, 1974), (Norman and Rumelhart, 1975).

A proposition is considered to be the smallest unit of meaning to which a truth value can be assigned. It consists of one predicate and arguments. Each argument is a concept extracted from the sentence, or it may be another proposition, and it fills a certain role for the predicate (*Agent, Object*, etc.)

For the current endeavour, the truth value of a proposition is not of interest. Truth value of an expression is a matter of debate and depends on the view one holds of language (mapping versus reality construction (Grace, 1987) or objective versus subjective versus experientialist (Lakoff and Johnson, 1980)).

What we are interested in is establishing connections between concepts expressed in the sentence through various syntactic structures. Because we focus on semi-technical texts, as explained in Chapter 1, the truth value of the sentences is not an issue.

Our approach emphasizes the role of relations between concepts. This view contrasts with the emphasis on discrete signs or symbols in Saussurian structural linguistics (de Saussure, 1959) and Chomskian generative grammar (Chomsky, 1965).

2.3 Theoretical Investigations of Semantic Relations

There are four leads as to how semantic relations became a research topic in linguistics.

1. The first one comes from probably the 5th century B.C.[3], from the work of Panini, a grammarian who analyzed Sanskrit, the language of Hindus at the time. In his main work, *Astadhyayi*, Panini identifies verbal and non-verbal relationships. In verbal relationships one of the elements must be a verb, the other a non-verb (Misra, 1966). The verbal relationships are further split into six *karakas* (literally *factors of action*). They correspond to six cases (the nominative, accusative, dative, instrumental, locative and ablative):

Apandana (*take off*): "(that which is) firm when departure (takes place)". In other words, it signifies a stationary object from which a movement proceeds. Equivalent to the ablative case. *Apandana*, the starting point, can be one of the following three kinds:

1. that in relation to which a movement is mentioned;
2. that in relation to which the verb expresses the movement only partly;
3. that in relation to which some movement is required.

Sampradana (*bestowal*): "he whom one aims at with the object". In other words, the recipient in an act of giving or similar. Equivalent to the dative case.

Karana (*instrument*): "that which effects most". Equivalent to the instrumental case. Whenever, after the activity of something, the action is meant to be conveyed as accomplished, that something is said to be the instrument.

Adhikarana (*location* or *substratum*). Equivalent to the locative case.

Karma (*deed/ object*): "what the agent seeks most to attain". Equivalent to the accusative case. *Karma* can be of three kinds:

1. product (*nirvartya*) – He made a jar out of mud.
2. modification/conversion (*vikarya*) – He converted wood to ashes.
3. destination (*prapya*) – He saw a tree. (the object does not change)

[3] The date varies from 5th to 7th century B.C. in various works.

Karta (*agent*): “he/ that which is independent in action”. Equivalent to the nominative case. The agent is the basis of all activities. (Scharfe, 1977)

The *karakas* relationships can be either explicit, in which case they are expressed by suffixation or compounding, or implicit, when they are marked by case-endings.

The non-verbal relationships are called *sesa* (literally *the rest* or *the extras*). A *sesa* relationship can be appositional, and it is marked by agreement for case, gender, person, conjugation, etc., or non-appositional.

Scharfe (1977) makes the following remarks:

> If the notion of *karakas* was perhaps derived from an observation of Sanskrit cases, Panini had raised them above the level of case values, and made them intermediaries between reality and the grammatical categories. Their importance, often misunderstood, goes far beyond the syntax of cases; next to the roots, they are the prime moving factors of the whole grammar.

2. Closer to our days, Lucien Tesnière in his work on syntax (Tesnière, 1959), follows a view, shared by many philosophers and linguists, that in addition to the logical/ propositional value of the sentence, the sentence is a mode of reflecting events in the world in a somewhat pictorial manner. For Tesnière, the meaningfulness of a sentence was due to the centralizing role of the predicate verb, which represented an action, and functioned as the highest syntactic node of the sentence. The verb is the complete and the independent term of a sentence.

The dependents of the verb are grouped in two classes.

1. **Actants**. They are the entities which participate in some way (even as mere onlookers) in a process described by the main verb. There can be at most three actants for any one verb:

   (a) *first actant.* From a semantic point of view it is the one performing the action. It corresponds to the grammatical subject.

   (b) *second actant.* It is the one undergoing/supporting the action. It corresponds to the direct object.

   (c) *third actant.* Semantically, it is the one who benefits or in the detriment of whom the action is performed. Syntactically it is the indirect object.

2. **Circumstances**. There are as many types of circumstances as there are of adverbs (time, place, manner, modality, degree, etc.).A verb can take a limited number of actants, but an unlimited number of circumstances.

Tesnière developed a representation of the structure of the sentence called *stemma*. The stemma representation for the sentence:

12 *John bought a computer for his son yesterday.*

is presented in Figure 2.1. *John* is the first actant (A1) involved in the *buy* event and the grammatical subject, *a computer* is the second actant (A2) and the direct object, *son* is the indirect object and the third actant (A3) or the recipient, and *yesterday* is a circumstance (C).

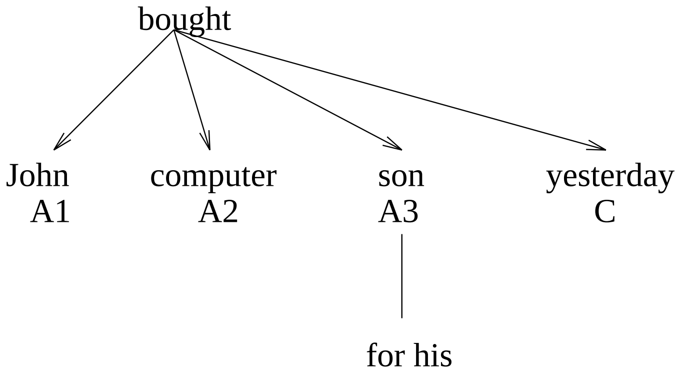


Figure 2.1: Stemmatic representation

This type of representation is different in content from a tree-diagram. While in the latter the connections between the nodes have no theoretical value, in the former such connections link the participants in an action. The stemmas transform the linear structure of the sentence into a diagrammatic representation of the meaning of the sentence conceived as an action.

3. As part of the research in syntactic theory done by Noam Chomsky and his group at the MIT in the 1960s, Gruber introduced six labels to describe the type of the entities in a sentence, relative to the main verb (Gruber, 1965):

Theme is the entity which is conceived as moving or undergoing transitions.

John gave a book to Mary.

Agent is the entity performing the action.

John bought a book from Bill.

Goal is the ultimate destination of the motion.

John ran below the deck.

Location indicates where the action takes place.

The bird was hovering nearby.

Accompaniment is the other participant in the action besides the Agent.

John flew the kite ahead of him.

Direction The essence of the expression of Direction is the specification of the path along which the Theme is traveling, but not to indicate the actual reaching of any Goal.

John ran toward the ocean.

Gruber further distinguishes two types of agent, depending on the meaning of the main verb:

- causative. The *Agent* is the one causing the action.
  John forced the wedge under the door.

- permissive. The *Agent* allows an action to occur.
  John let Bill watch television.

4. Also starting from a grammatical analysis of cases, Fillmore's analysis veers off into semantics and leads him to the definition of a set of relations between the main verb of the clause and its arguments. Fillmore starts from the analysis of prepositions in the framework of transformational grammar (Somers, 1987). In (Fillmore, 1968) he identifies a list of "needed" cases, which we quote together with the corresponding definitions:

Agentive (A), the case of the typically animate perceived instigator of the action identified by the verb.

Instrumental (I), the case of the inanimate force or object causally involved in the action or state identified by the verb.

Dative (D), the case of the animate being affected by the state or action identified by the verb.

Factitive (F), the case of the object or being resulting from the action or state identified by the verb, or understood as a part of the meaning of the verb.

Locative (L), the case which identifies the location or spatial orientation of the state or action identified by the verb.

Objective (O), the semantically most neutral case, the case of anything representable by a noun whose role in the action or state identified by the verb is identified by the semantic interpretation of the verb itself; conceivably the concept should be limited to things which are affected by the action or state identified by the verb. The term is not to be confused with the notion of direct object, nor with the name of the surface case synonymous with accusative.

Fillmore does recognize that this is not a comprehensive list. Further in the article he mentions a **Benefactive** case, and the possibility of the existence of others.

Fillmore calls these relations **cases**. They reflect the relations between the markers associated in Indo-European languages (for example Romance languages) and the arguments of the main verb. In English most of these markers have disappeared, but the role the arguments play in the semantics of the clause remains.

Fillmore proposed that these cases be regarded as universal relations, whose function is to designate the possible semantic relationships between the main verb of the sentence and its nominal arguments. He also proposed the replacement of the deep syntactic structure (advocated by Chomsky) with a case structure. Despite the semantic side of the postulated case structure, the type of evidence used to support their introduction in theory is surface-syntactic. They are used to explain grammatical phenomena, such as the selection of the subject, well-formedness of a sentence, etc.

Modern theory of semantic relations stems from Fillmore's work ((Fillmore, 1977), (Jackendoff, 1972), (Jackendoff, 1976)). The cases in the sense defined by Fillmore have been renamed thematic roles, or θ-roles.

The set of thematic roles associated with a verb is sometimes called a **θ-grid** or a **subcategorization frame**. The theory associated with thematic roles is based on the following axiom:

> There must be a one-to-one mapping between θ-roles and arguments of a verb.

In the **case** theory of relations between the verb and its arguments, cases were assigned only to the arguments that a verb subcategorized for. Some arguments of the verb are considered adjuncts, and not part of the subcategorization frame, as in the examples:

13 *He walked 4 miles.*

4 miles is a measure of the action, but it is not subcategorized for.

Choosing only the required arguments for a predicate restricted the list of possible relations. The theory was extended to include all possible arguments.

The list of θ-roles is still a much debated issue. Its length varies from a few to hundreds of labels. Dowty (1991) shows that there is the problem of role fragmentation, as illustrated by different refinements of the *agent* role in the literature: *agent* and *actor* (Jackendoff, 1983); *agent* and *effector* (van Valin, 1990); *volitive, effective, initiative* and *agentive* (Cruse, 1973); fourteen different groups proposed by Lakoff (1970). Gruber (1965) also identified two possible types of *agent*: *causative* and *permissive. Theme* can also be divided into *incremental, non-incremental* and *holistic themes.*

Dowty then proposes a different view, in which these roles are not discrete categories, but fuzzy concepts that can be better described by prototypes. He introduces the concepts of *Proto-Agent* and *Proto-Patient*, characterized by the following lists of features.

Agent Proto-role

- has volitional involvement in the event or state;
- has sentience (and/or perception);
- causes an event or change of state in another participant;
- moves (relative to the position of another participant);
- exists independently of the event named by the verb.

Patient Proto-role

- undergoes change of state;
- is an incremental theme;
- is causally affected by another participant;
- is stationary relative to movement of another participant;
- does not exist independently of the event.

According to the number of features which they share with these proto-roles, the arguments of the verb will be more agent- or more patient-like. The weakness of this theory is the origin and the nature of the features relevant to proto-role assignments (Levin and Rappaport-Hovav, 1996).

Chomsky's work on deep and surface structures as part of his syntactic theory has spilled from verb and its arguments to nouns and their modifiers, transferring also the roles of a verb to its nominalized version (Chomsky, 1970). There is a difference between the occurrence evoked by a verb and the occurrence evoked by the verb's nominalized version. The nominalized form usually refers to the event described as a whole, whereas the verb shows an action that is unfolding. Such differences can be marked in a semantic description of the sentence, and we support the view that the same label describes the interaction between the head of the phrase under analysis and its arguments (Quirk et al., 1985).

Relations between acts as a whole, as opposed to relations between an act and its arguments, were first introduced by Schank (1975). In his conceptual dependency representation he introduces relations between such acts as *transfer (ATRANS - abstract transfer)*. For example, in the analysis of the sentence:

14 *John bought a car from Bill.*

there are two *ATRANS* acts, as shown in Figure 2.2, one to transfer the possession of the car from Bill to John, and the other implied by the verb *bought* to transfer the possession of money from John to Bill. The relation between these acts is Instrument.

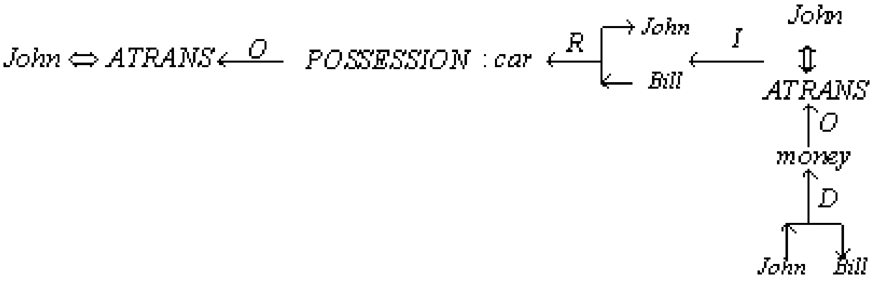


Figure 2.2: Conceptual representation using Schank's primitives.

This theory was later refined in Schank and Abelson (1977). These relations do not stem from an analysis of syntactically related clauses, but from simple clauses whose elements, when represented through primitives, show an interaction between acts.

In analyzing relations between clauses, two types of relations are considered: semantic and pragmatic. Rhetorical structure theory covers both these aspects (Mann and Thompson, 1986a), (Mann and Thompson, 1986b), (Mann and Thompson, 1988). These relations raise above the level of the sentence and explore relations between larger text units.

Such discourse relations were explored by Hovy (1993), and indicators for the appropriateness of one label over another were also found. However, relations between sentences are not the focus of this dissertation. Clauses that can be combined to form multi-clausal sentences will be addressed by the inter-clausal relations that we explore. Clauses that cannot are considered to be too far apart with respect to semantic relations. They may be connectable through a chain of relations, where most of the intermediate information is implicit, as in the following example:

15 *We were late for the party. The engine broke down.*

There is a causality chain that involves recognizing the fact that we were going to the party by car, whose engine broke down, causing us in the end to arrive late. While we do not exclude this type of relations from our research, we do not address them at this point. Resources that allow a system to automatically infer connections between such sentences are being built (Lenat, 1995),(Singh et al., 2002), but are not yet available. One other reason for not dealing with these relations is that we use structure to reach the meaning. There are no formal accounts of the structure of a text above the level of the sentence, that would allow us to pick clues – in the form of connectives for example – about the underlying relation.

Barker (1998) presents a comprehensive study of the research on semantic relations from the point of view of text analysis. This study presents the various lists of semantic relations proposed by researchers in text analysis. We observe throughout the lists of relations presented therein that research is focused on syntactic units pertaining to specific syntactic levels: a verb and its arguments, a noun and its modifiers, or clauses.

Apart from research that focuses on connecting pairs of syntactic units through semantic relations, there is a large body of work that *uses* semantic relations. Research on representing knowledge extracted from texts in machine-readable format uses semantic relations to label the links between the lexical concepts extracted from texts. Sowa (1984) has started a current of research in this direction by proposing a conceptual graph notation for the analysis of texts. The analysis relies on representations of concepts as a collection of primitives.

For example, ANIMAL is [ANIMATE,MOBILE-ENTITY,PHYSOBJ,¬MACHINE][4] (Sowa, 1984, p. 408), BELIEVE is [STATE]

Semantic relations are defined based on the primitives of the concepts they connect. For example, the relation (DUR) (duration) used in conceptual graph representation, is defined as "(DUR) links a [STATE] to a [TIME-PERIOD], during which the state persists." (Sowa, 1984, p. 416). *5 hours* is an example of time period.

We define semantic relations as connecting concepts – occurrences, entities and attributes. It is interesting to observe here a parallel with the definitions presented in (Sowa, 1984): various primitives are used to explain what the semantic relations defined connect. We use occurrences, entities and attributes, which at a very general level can be considered as being primitives as well. The difference comes from the fact that we choose to view all concepts as occurrences, entities and attributes with a specific purpose in mind: this coarse-grained view allows us to bridge structures in syntax like clauses and nominal, adjectival and adverbial phrases with concepts.

Since we aim to find bridges between concepts and the syntactic structures that bring them forth, research that uses semantic relations as a tool that serves a specific purpose, without investigating them, does not bring new information. It is interesting for us, for example, to find all possible expressions of temporal indicators. A time interval is not only an explicit temporal expression like *5 hours*. Occurrences have an implicit time frame, which can be used as references for other occurrences. The temporal or spatial dimension of occurrences is not part of the description in terms of primitives of lexical concepts. Words in certain contexts acquire a different meaning than the one which is described in a lexical resource through primitives. EXAM for example may be described as [WRITTEN TEST]. But in the sentence:

16 *The student was anxious because of the exam.*

exam represents an occurrence that involves *writing, answering* or another action connected to exams that causes students to be anxious.

Defining semantic relations based on characteristics of the concepts they connect is an interesting aspect. We explore it in Chapter 6. We will use ontologies of concepts to find what concepts connected by the same relation have in common.

[4] Conceptual primitives are shown in square brackets.

Since in research on knowledge representation semantic relations are not the object of investigations, but tools that serve a specific purpose, we do not present such research here.

2.4 Present Status

Present-day research is not much concerned with general issues concerning semantic relations. It focuses mostly on specific domains..

Systems are designed to analyze texts in a certain field. They use lists of semantic relations specifically tailored to capture salient connections between concepts in the domain. Sometimes they also use lexical resources developed to describe the concepts in the field in question.

One of the best known examples is the FrameNet project at the University of Berkeley, California (Fillmore and Atkins, 1998),(Baker et al., 1998) (Gildea and Jurafsky, 2002). It proposes case frames used to analyze texts pertaining to law. The project leader is Charles Fillmore. The case frames developed label each participant in a specific type of legal event. The participants are extracted at the intra-clause level as arguments of the verb. For example, the *Criminal_process* frame includes reference to a *Suspect* which has been arrested by an *Authority*, and against which are pressed *Charges*.

Another project focused on a specific domain is BioText, also at the University of Berkeley, CA under the leadership of Marti Hearst (Rosario and Hearst, 2001),(Rosario et al., 2002). The project aims at identifying entities and relations between entities in bioscience texts. The authors make use of an ontology of concepts built from medical texts – MeSH (Medical Subject Headings). The relations on which they mostly focus are relations between components of nominal compounds. Rosario and Hearst (2001) propose 38 relations, more specific than the generic Agent, Object, etc., but specific enough to be useful for their task, for example *activity/physical process* (virus reproduction), *change* (disease development), *cause (1-2)* (food infection), *cause (2-1)* (flu virus), *defect* (hormone deficiency), *procedure* (blood culture), etc.

SNOWY is a knowledge acquisition project developed by Fernando Gomez at the University of Central Florida (Gomez, 1998b) that processes general texts. The semantic interpretation part of the project is based on a list of thematic roles, an ontology of predicates connected to *WordNet*'s verb classes, and connections between these predicates and *WordNet*'s ontology of nouns. The thematic roles (*agent, theme, instrument*, etc.) apply to links between verbs and their arguments, and also to nominalized verbs and their modifiers.

Rapid Knowledge Formation (RKF) is a recently concluded project at the University of Texas at Austin whose goal was developing a system for building complex knowledge bases through the combination of components (events, entities and modifiers) (Clark and Porter, 1997). For the description of the relations between these components, the project makes use of a dictionary of relations that describe the interaction between two events (e.g. causal relations), an event and the entities involved (e.g. *agent, instrument*), an entity and an event (e.g. *capability*) two entities (e.g. *part*) or an event or entity and their properties (e.g. *duration, size*) (Fan et al., 2001). These relations cover three syntactic levels and stem from (Barker, 1998). The system is used by experts in a certain domain to encode the knowledge in their specific field.

2.5 Conclusions

Semantic relations have evolved from the desire to get a grasp on the meaning behind a sentence, by understanding the roles played by the entities involved in an event. The notions of *subject, object, indirect object* provided by grammars do not offer a good enough explanation. *Subjects* for example may denote entities actively involved in an action, just passive onlookers, or the ones affected by the event. The traditional notion of case has been the starting point in the development of a set of roles that described the involvement of different entities in the situation captured in a particular sentence. We have seen that the idea of semantic relations, initially based on grammatical cases, has appeared, independently, at four different points in time. From these theories, in particular mostly Fillmore's theory of thematic roles (Fillmore, 1968), the idea of semantic relations has expanded first to all arguments of the main verb, not just the ones it subcategorizes for, and afterward to noun phrases, through nominalizations, to clauses and beyond the borders of the sentence to larger units of text.

Throughout the literature we observe two main views on semantic relations.

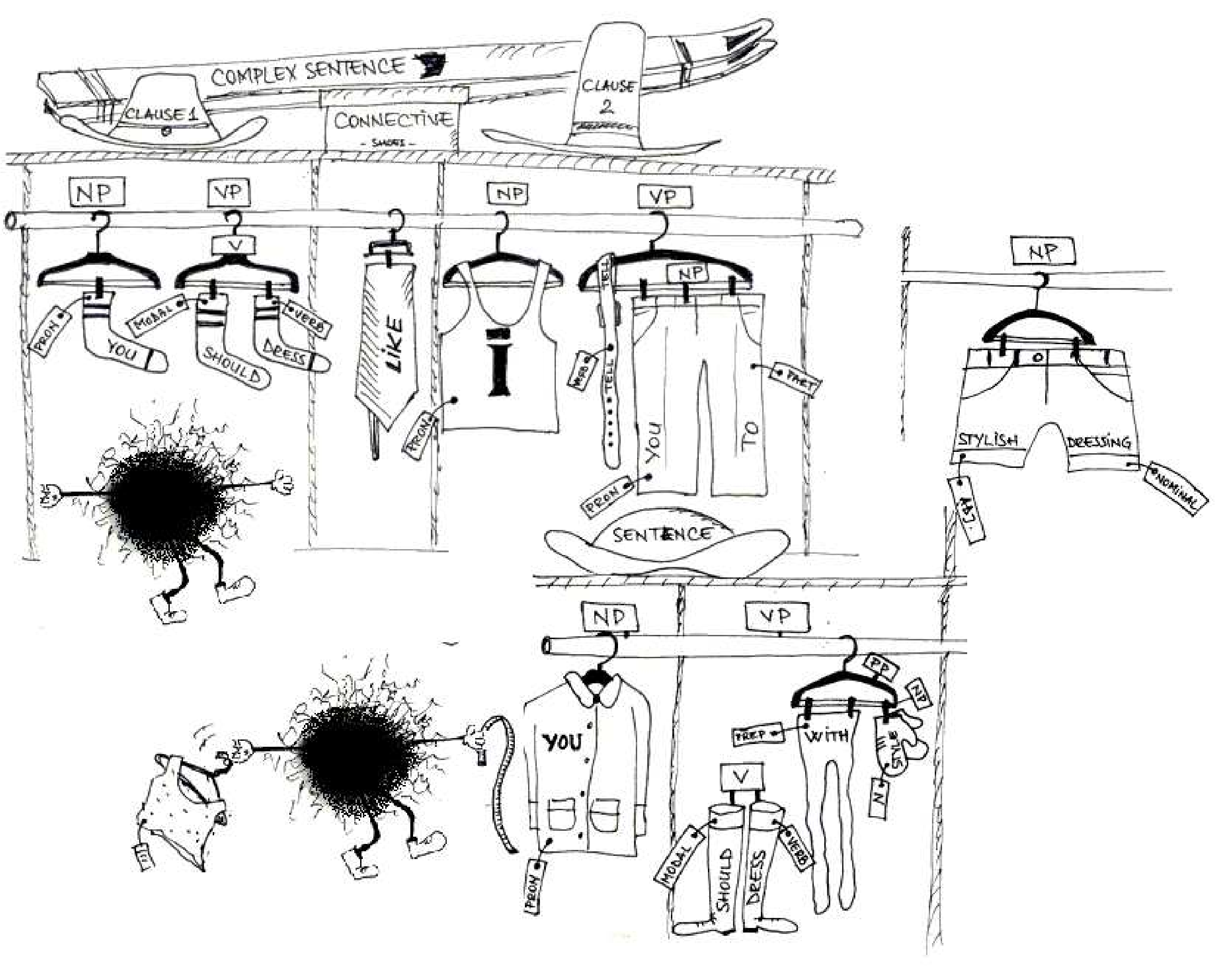
One pervasive view is that semantic relations describe the interaction between syntactic units. This is the view rooted in analyzing semantic relations starting from grammatical cases, as seen in the works of Panini (Scharfe, 1977), Tesnière (1959), Gruber (1965), Fillmore (1968) and their followers. Through the brief overview of current endeavours in semantic analysis using semantic relations presented in Section 2.4, we show that the view that semantic relations hold between syntactic units is still dominant in this type of research. The FrameNet project focuses on the intra-clause while BioText on the noun-phrase level. SNOWY deals both with the intra-clause and the noun-phrase level. The connection between the two levels comes from dealing with nominalized forms of verbs. The same relations that apply at the intra-clause level apply at the noun-phrase level. The RKF project did not involve automatic text processing, but the relations they assigned as part of the knowledge formation process pertained to three syntactic levels.

The other view comes from researchers in whose work semantic relations are used for text analysis, namely assigned to pairs of words extracted from texts, as is the case, for example, in building conceptual graphs (Sowa, 1984). The semantic relations in this case do link concepts. Those are *lexical concepts*, where the meaning of each word is analyzed in terms of features, or according to other theories of concepts like the ones we present in Section 3.2.

By anchoring semantic relations into concepts like occurrences, entities and attributes, as opposed to syntactic units or lexical concepts, we can find justifications why, for example, a semantic relation between a noun and its modifier is the same as the relation between two clauses. In the work reviewed such assumptions may have been made, but not justified. We will give systematic justifications why semantic relations are the same across syntactic levels. We find such justifications by explicitly adopting the view that semantic relations link occurrences, entities and attributes. If an occurrence, for example, can surface in various forms, then these forms are responsible for instances of the same semantic relations on various syntactic levels. We will investigate in Chapter 3 whether the same, or very similar, concepts can surface in various forms.

Chapter 3

Surface Forms of Concepts



3.1 Introduction

We want to provide systematic evidence that semantic relations are the same across syntactic levels. To do this we look at what semantic relations are, what they mean and what it means for a semantic relation to be the same on different syntactic levels.

We have already established that semantic relations can be seen to describe connections between two concepts. The nature of the semantic relation and the nature of the concepts they connect impose certain constraints on each other. For example, causality relations require that the concepts linked be occurrences. Conversely, if a causal relation is assigned to a pair of concepts, these concepts are occurrences, regardless of their syntactic expression[1].

We use this coarse view of concepts, which distinguishes only between occurrences, entities and attributes of occurrences and entities, in order to find parallels between an idea that we want to convey, and the linguistic expression that conveys it. Linguistic expressions are structured, and the structure is available to us via grammatical analysis. The type of structures that grammars provide – clauses, nominal, adjectival and adverbial phrases[2] – are the ones that bring forth occurrences, entities and attributes. If occurrences, for example, can be conveyed through various types of syntactic structures – a clause, or just a noun-phrase – this shows that a semantic relation can have instances on different syntactic levels. The syntactic level on which it appears is dictated by the level on which the concepts it connects appear. Then, since the concepts that surface in various forms are actually the same, the relation that connects them is the same, regardless of the syntactic level on which the concepts surface. This means that semantic relations are the same across syntactic levels.

Throughout this chapter we focus on exploring the idea that a concept can surface in language through different expressions, and we will look for explanations that relate the various surface forms that a concept can take.

In Section 3.2 we present and explain the term *concepts*, as we use it throughout this dissertation. We compare the definition we adopt with other definitions in the literature, and we show that the view we adopt has no negative impact on any of the theories of concepts that we present.

Section 3.3 includes arguments for and against the idea that the same concept, or

[1] Not all researchers share the view that causality relations hold between two occurrences. This will be discussed in Chapter 4.

[2] We focus on structures whose head is an open-class word, for reasons presented in Section 3.4.1.

type of concept, can surface through different expressions. We explain our view on this problem, and show what we look for when we talk about surface forms of concepts: ways to map occurrences, entities, and their attributes onto various syntactic structures. We then proceed to investigate the way in which humans produce language, to search for explanations why the same concept can have alternative expressions. We look for clues in various fields concerned with language production: psycholinguistics, language generation, generative grammars and lexicology.

In Section 3.4 we proceed to find our own account of phenomena that relate different surface forms of the same concept. We first establish the forms that concepts can take, by grouping syntactic units inside a sentence into structures centered on open-class words. We then investigate how a concept – occurrence, entity or attribute – can be mapped onto each of the members of every possible pair of such structures, and then we identify the phenomena that relate them. Once we have investigated all possible mappings and justified those that are not possible, we take each phenomenon identified, and we discuss its impact in light of literature review and examples. These phenomena that we identify are not particular to certain concepts, but provide a general explanation of why occurrences, entities and attributes can take different surface forms.

Section 3.5 explains why the phenomena identified and discussed in Section 3.4 are relevant, and how we use them in the remainder of the dissertation.

In Section 3.6 we investigate briefly other languages in the Indo-European family, to show that the phenomena we have identified are not particular to English. We also give a few examples in some more exotic languages. We propose a more thorough analysis of other languages for future work.

3.2 Concepts

Jackendoff (1989, p. 305) says:

> Asking a psychologist, a philosopher, or a linguist what a concept is is much like asking a physicist what mass is. An answer cannot be given in isolation. Rather, the term plays a certain role in a larger world view that includes the nature of language, of meaning, and of mind. Hence the notion of a concept cannot be explicated without at the same time sketching the background against which it is set; and the "correctness" of a particular notion of concept cannot be evaluated without at the same time evaluating the world view in which it

plays a role.

Our purpose is to find a bridge between the concepts in our mind and the linguistic expressions that bring them forth. From this perspective, we define concepts to be the type of things that we talk about: occurrences (all types of processes, events, activities, states, actions, accomplishments), entities (concrete and abstract objects), and attributes of occurrences and entities. By adopting this definition, we can map concepts onto the syntactic structures of linguistic expressions. We have occurrences, entities, attributes of occurrences and attributes of entities on one side, and verb-centered structures, noun-centered structures, adverb-centered structures and adjective-centered structures[3] on the other.

Occurrences are not only what we express in language through verbs, but the entire situation, including its participants and attributes. According to this definition, occurrences, as opposed to entities and attributes, seem to be complex concepts, in the sense that they include other concepts that describe their participants and their attributes. This is true, although not entirely. We adopt a holistic view of occurrences, in which we do not separate the verb that describes the action/event/state/process from the nouns/adjectives/adverbs that describe the concepts involved in the occurrence. We do single out the concepts that describe the participants when we talk about relations between an occurrence and the entities involved. The occurrence, however, is still seen as encompassing them all.

Despite this clarification, the investigation that we propose of possible surface forms of concepts and relations between them does not assume that concepts are atomic.

The term *concept* is very loaded, especially in cognitive circles. There are several theories of what concepts are. We will mention briefly each of these theories, and then we will explain that the view that we adopt does not infirm, confirm or in any way affect any of these theories. The term *concept*, in the sense that we use it according to the definition above, can be seen as a *meta concept*, in the sense that it proposes a view that groups concepts, as seen by cognitivists, into 4 major classes: occurrences, entities, attributes of occurrences and attributes of entities. This has no impact on the "fuzziness", "prototypicality", and other aspects of concepts that these theories propose.

Margolis and Laurence (2002) show a comprehensive review of theories about *concepts*, "the most fundamental constructs in the theories of mind" [pg.3]. The collected articles

[3] X-centered structure simply means the syntactic structure whose head is a word that has X part of speech.

show how theories of concepts have appeared and changed, following philosophical, psychological and cognitive explorations into the nature of our mind. Each theory has its strengths and weaknesses, but not one of them can by itself explain all the questions about concepts and the way we use them.

We present the five theories discussed in (Margolis and Laurence, 2002), to compare the views of concepts they propose, with the view we adopt in this research.

The Classical Theory of Concepts According to this theory, “most concepts (esp. lexical concepts) are structured mental representations that encode a set of necessary and sufficient conditions for their application, if possible, in sensory or perceptual terms.” (Laurence and Margolis, 2002, p. 10) The oldest and most enduring of concept theories, the classical approach has roots in antiquity in the works of Plato. The reason it has not been challenged until very recently (1950s) is that its intentional view of concepts offers valid explanations for concept acquisition, categorization, epistemic justification, analytic entailment and reference determination.

The Prototype Theory of Concepts “Most concepts (esp. lexical concepts) are structured mental representations that encode the properties that objects in their extension tend to possess.” (ibid., p. 31)

The Theory-Theory of Concepts “Concepts are representations whose structure consists in their relations to other concepts as specified by a mental theory.” (ibid., p. 47)

The Neoclassical Theory “Most concepts (esp. lexical concepts) are structured mental representations that encode partial definitions, i.e., necessary conditions for their application.” (ibid., p. 54)

Concept Atomism “Lexical concepts are primitive; they have no structure.” (ibid., p. 62)

Lexical concepts are considered to be concepts like BACHELOR, BIRD, BITE – concepts that correspond to single morphemes in natural languages (ibid., p. 4).

For us concepts do not correspond to single words, but to sequences of words of any length. For the purpose of finding correlations between concepts and the syntactic structure through which they surface in language, we defined concepts based on the type of things that we talk about – occurrences, entities and their attributes – and to which we can map syntactic structures. We do not constrain these type of concepts to be atomic or not – this has no relevance to our study.

The notion of concept that we propose is more general than the ones defined in these theories. For the purpose of the theoretical discussion that follows it is only interesting for us to know whether the concept is an entity (a concrete or abstract object), an occurrence or an attribute, and not to have a definition or representation of the concept in terms of primitives, features, etc. Therefore the definition of concept does not commit to any of the theories presented, nor does it provide evidence for or against any of them.

In this chapter we will first defend the idea that the same concept can surface in various forms. We will see that this idea has strong opponents. However, we adopt a more simplistic view of the meaning behind a linguistic expression, which allows us to concentrate on the concepts expressed and the links between them. We focus on the part of meaning which consists of the concepts conveyed and the way they interact. For the purpose of this study we do not take into account issues like focus, quantification, mode and tense, which all introduce variations in the meaning of a message. We recognize their importance, however, and we plan to incorporate them back in our study, as stated in Chapter 7. Also, because of the fact that for us concepts do not represent fine-grained *lexical concepts* but rather categories (occurrences, entities and attributes), the issue whether a concept can surface in different forms is not sensitive to nuances in meaning.

Once we have shown that from our perspective of the message conveyed, concepts can surface through various expressions, we will investigate syntactic evidence for this fact, and we will analyze phenomena that relate different surface forms of the same concept[4]. Concepts are not only what underlies a certain word or a sequence of words, but everything that the word/sequence of words stands for, in the context given by the text under analysis. Let us compare for example the word *sweets* and the concept it represents in isolation, as opposed to the concept implied when it appears in the context:

17 *Sweets before dinner spoil your appetite.*

In this case, *sweets* does not only represent the concept of SWEET EDIBLE THINGS, but the idea of EATING these things.

The idea underlying sentence (17) can be equivalently expressed as:

18 *Eating sweets before [eating] dinner spoils you appetite.*

[4] The issue whether the same concept can surface through different expressions is discussed in Section 3.3

What surfaces in the syntactic expression of an idea is a projection of concepts onto words that are meant to evoke a similar idea in the mind of the listener. We cannot say that it will be the same idea, just that it is very similar, because the same words can have different connotations for different people.

Example (17) illustrates one type of phenomena that accounts for different surface forms for the same concept. In this situation, a word conjures up a whole occurrence in the mind of the reader. There are cases in which it is not the same concept that surfaces in different forms, but the same type of concept. For example, a temporal or locative expression can be explicitly (*at 5 o'clock, near the statue*) or implicitly conveyed through an occurrence that expresses an equivalent type of time (punctual, bounded or unbounded interval) or spatial indicator (*when you arrive, where we met last time*).

In order to identify phenomena that allow for a concept to take different surface forms, we work on the evidence found in texts, some available lexical resources, and the semantic relations mapped onto a network of words. We identify the concepts that underlie the surface form. Semantic relations are identified, named and assigned by humans. The knowledge people bring into this process goes beyond anything that any present language analysis system has available. We want to take advantage of this knowledge, and use it to detect elements of text and evidence of syntactic changes that indicate the presence of specific types of concepts that are not there explicitly, or only partially there.

The way in which a thought finds a linguistic expression is still a mystery.

First of all, the thought in its pre-verbal form is considered not to be available to the conscious mind. It exists only at a subconscious level (Jackendoff, 1994). The moment it becomes conscious, it has already been verbalized:

> ... the language we hear in our heads while thinking is a conscious manifestation of the thought – not the thought itself, which isn't present to consciousness.

Grace (1987, p. 10), while presenting his reality-construction view of language, says:

> It is impossible to draw a clear line between thinking, i.e. bringing a thought into being, and encoding a thought, i.e. putting it into words.

It is very tempting to speculate how thoughts are actually represented in the mind (Fodor, 1975). It is hard to provide evidence for such views.

We are interested in the ways in which a concept can surface in language. We will judge the similarity of different forms that a concept can take with respect to the semantic relations in which it is involved.

We will show that a concept can surface as a verb-centered structure (or clause), as a noun-centered structure (noun phrase), or even as a single word. Different forms have different implications. Using a clause is equivalent to asserting that the occurrence described took place. The same cannot be said about a noun phrase, for example. The difference between the two is the difference between asserting and saying (Grace, 1987).

Grace (ibid.) proposes the view that the content of a linguistic expression is given by the specification of a conceptual event (specification of an abstract event or situation), plus contextualization clues (indications as to how this concept fits in the ongoing discourse), plus modality (conditions of instantiation – assertion, interrogative, negative, contingency of various sorts (will, may, etc.)). In this work on semantic relations, we concentrate only on the specification of a conceptual event – its linguistic form. From this point of view, the fact that an utterance is asserted or not does not make a difference. We consider whether the underlying concepts are the same, and whether they are linked by the same semantic relation.

3.3 Saying the Same Thing

3.3.1 Can We Say the Same Thing in Different Ways?

Can we express the same ideas in different ways? This is equivalent to asking whether different syntactic expressions can be used to convey the same meaning.

There are researchers who claim we can. It is the case of transformationalist grammarians, for example, who claim that the same deep structure that captures the meaning can have different surface – syntactic – forms (Chomsky, 1966). A simple example would be the active-passive alternation. According to the transformationalist paradigm, the following two sentences convey the same meaning:

19 *John sold the car to Bill.*

20 *The car was sold to Bill by John.*

Research on paraphrasing holds a similar view. In computational circles, research in paraphrasing concentrates on finding alternative expressions of the same idea using different translations of the same text for example (Barzilay and McKeown, 2001), (Barzilay and Lee, 2003), using different news reports on the same topic (Shinyama et al., 2002), building paraphrases that comply to certain constraints (like length for example) (Dras, 1997), etc.

At the other end of the spectrum are researchers who hold the view that if there are two distinct forms in language, they must convey different meanings, even if the difference is very fine. Alternative forms of expression have evolved and remained in language because existing forms did not convey the meaning intended by the speaker (Bolinger, 1977), (Dowty, 2002). The two sentences presented above would differ in emphasis. In sentence (19), the emphasis was on the *Agent*, the one performing the *sell* act. In sentence (20), the emphasis is on the *Object* of the transaction. Other syntactic variations have special implications for the occurrences described. For example:

21 *They loaded the truck with hay.*

implies that the truck is full of hay, whereas the sentence:

22 *They loaded hay onto the truck.*

makes no such implication.

We acknowledge that different expressions are different in meaning, and sometimes this difference is very fine. For any pair of expressions it is easy to explain the difference. We want to focus instead on similarities. If two expressions with different syntactic structure, E_1 and E_2, are involved in the same semantic relation with the same expression E, we want to see what E_1 and E_2 have in common, even if their meaning is not the same.

To illustrate this point let us consider two examples:

23 *I will leave at 5 o'clock.*

24 *I will leave when you arrive.*

The two temporal expressions *5 o'clock* and *you arrive* mean two completely different things. However, they have something in common: they both fix a point in time. *5 o'clock* indicates a specific point on the time axis, while *you arrive* is a punctual occurrence that will happen at a point on the time axis. The two expressions, then, convey the same type of concept TIME POINT. Because of this similarity, both these expressions can serve as a reference on the time axis for the occurrence conveyed by the expression *I will arrive.*

The situation captured in the examples (25) and (26) is different. In this case different expressions convey the same concept, not type of concept, while taken outside their context they mean completely different things.

25 *Drinking coffee after 5 o'clock doesn't let you sleep.*

26 *Coffee after 5 o'clock doesn't let you sleep.*

In the context in which it appears *coffee* represents the concept of DRINKING DARK LIQUID MADE OF COFFEE BEANS, just as *drinking coffee* does.

The same word or expression used in different contexts can convey a different message. But what happens with different expressions in the same context? This is what we are interested in, whether expressions used in the same context can convey the same message (from the point of view of the concepts or types of concepts conveyed), as we have seen in the examples (23) through (26). The context in which we judge different surface forms is the expression with which they are connected, and the semantic relations that describes this connection.

When we talk about surface forms of concepts, what we look for are alternative expressions of occurrences, entities and their attributes, to which correspond different types of syntactic structures: verb-, noun-, adjective- and adverb-centered structures.

Throughout our discussion we will present examples in which we try to keep the underlying message the same or as similar as possible. The reader should be able to compare the different surface forms in the light of the message conveyed, and the semantic relation under scrutiny in each particular instance.

Syntactic forms can vary on a lexical or a structural dimension. Lexical variation (for example using synonyms) does not change the structure of the sentence, therefore the relations among concepts will remain on the same level. We are interested in linguistic phenomena that account for expressing the same idea using different structures, so that relations that were on one syntactic level now appear on a different level.

We will look at different fields of research that are concerned with the way utterances are produced. Research in psycholinguistics is interested, among other things, in discovering the processes that unfold in the mind of the speaker while producing an utterance, or in the mind of the listener while understanding it. Language generation concentrates on producing sentences or larger texts. Part of the generation process is inspired by psycholinguistic research in language production. We seek in these fields clues and rules for producing different syntactic expressions for the same underlying ideas. Transformational grammars claim to give an account for different surface structures for the same deep structure. We will look into that as well, and see what phenomena are covered by transformational rules. The last field we will look at is lexicology. The phenomenon we are interested in is word formation, in which words are coined to replace more complex expressions.

We then try to synthesize some observations into phenomena that justify why the same idea can be mapped onto different syntactic structures. We give a comprehensive account of the phenomena identified. *The same idea* means, for us, the same concepts connected in the same way. "The same idea", as explained above, is relative to the context.

3.3.2 Language Production

Language production studies make the premise that there is something, a message, that the speaker wants to convey, but will not attempt to show how this message is represented before it assumes a linguistic form (Fromkin and Ratner, 1998). Research concentrates on the elements into which the conceptual message is translated, and how these elements are combined in order to convey it.

As mentioned before, all these processes are subconscious. In order to gain insight into their nature, researchers use instances when the workings of the subconscious reveal themselves. This happens when the speaker produces different types of "slips of the tongue" or disfluencies in speech. It is considered that in such cases, the subconscious takes over (Freud, 1901).

Insight into the ways in which humans produce speech comes also from the analysis of patients with brain lesions that affect areas of the brain used in speech production and comprehension (Butterworth, 1980a).

Analyzing errors in speech gives a glimpse into the transformation from pure thought to verbalized thought. Such errors may occur at different levels: the level of unit of speech (phonemic segments, phonetic features, syllables, stress), the level of word, morpheme or phrase unit (word selection and placement errors, lexical search and pausal phenomena, morphemes and speech errors, grammatical rules), planning level (phrase reversal, self-corrections and retracing, pausal phenomena) (Bock and Levelt, 1994).

There are four major models of utterance generation that explain each of the stages involved in producing a linguistic form for a conceptual message. Decoding the processes pertaining to each stage of linguistic form production is based on results of the research into the nature of the production process using errors detected in speech.

Utterance generator model (Fromkin, 1971). It proposes the following stages:

1. **Generate a meaning to be conveyed**. The conceptual message is subcon-

scious.

2. **Map a message onto a syntactic structure**. A syntactic skeleton is created, which will be fleshed out with words in a later stage. This structure determines the form and grammatical category of words that may be chosen.
3. **Generate the sentence and phrasal stress**.
4. **Select words from the lexicon**. This process is guided by the syntactic structure, and the constraints on word categories encoded in this structure.
5. **Apply phonological pronunciation rules**.
6. **Generate the motor commands for speech**.

The Garrett model (Garrett, 1980), (Garrett, 1984). It can roughly be split into the following stages:

1. **Obtain a message representation using inferential processes**.
2. **Functional processing**. This includes lexical selection and determination of the functional structures. The functional level is "a multiphrasal level of planning in which the assignment of major lexical-class items to phrasal roles is accomplished". The result of this stage consists of a functional representation of the original message.
3. **Positional processing**. It involves the retrieval of word forms, assignment of segmental and prosodic structure for words, assembly of the constituents. The result of this stage consists of a positional representation.
4. **Phonological encoding**. A sound-level representation is produced.
5. **Send instructions to articulators**.

Levelt's model (Levelt, 1989), (Bock and Levelt, 1994). This model has the following components:

1. **Message generator**. A *preverbal message* is generated by conceptualizing the utterance.
2. **Formulator**. The preverbal message is grammatically (lexical items are retrieved) and phonologically encoded (the syntactic outline generated in the previous step is used to generate a phonological plan for the utterance. It includes intonation and stress patterns).
3. **Articulator**. The phonetic plan of the utterance is executed by sending instructions to the neuro-muscular system.

Working in parallel with this speech production model, Levelt proposes a *speech-comprehension system.* The role of this second system is to monitor the utterances produced for errors, in order to correct them.

Dell's model (Dell, 1986). The previous models proposed modular approaches to language production. Dell proposes a connectionist model for lexical organization and retrieval. The units of the networks are words, possibly also rules, and the connections between them are based on semantic and phonological relatedness. The model works by activation spreading over the network. This model does not account for other parts of speech production, other than lexical access.

Although language production is a subconscious process, psycholinguistic research has identified production stages through analysis of "slips of the tongue", and speech production in people who have suffered specific types of brain lesions. Errors have shown that there are several competing plans for producing a sentence to convey a certain message, both at the sentence structure and the lexical level (Butterworth, 1980c). Hypotheses about how these plans are built, and how they compete among themselves to determine the expression that is finally chosen come only from the connectionists (Dell, 1986),(Rumelhart and McClelland, 1986),(Tabor and Tanenhaus, 1999),(Dell et al., 1999), who propose that a message that the person wants to convey leads to the activation of connected concepts, and the words that can be associated with these concepts. The words that are more strongly activated are the ones that will be produced. Since the connections between words are based both on semantic and phonological similarity, as Dell's model explains, certain errors in speech can occur. Symbolic rules for language production are still a research topic (Bock and Levelt, 1994).

A more interesting idea, for our purpose of finding different expressions of concepts in language, comes from cognitive research. Analysis into the use of figures of speech – metaphor, metonymy, etc. – show how we use language, and more than that, how we conceptualize the world around us and how do we think in terms of the concepts we have formed.

Figures of speech are a fundamental part of our conceptual system (Lakoff and Johnson, 1980),(Gibbs, 1994). They allow us to extend the way we talk about things.

Metaphors, for example, are used to map concepts we are familiar with onto other concepts, based on analogy and perceived similarities. For example, we conceptualize time and space in similar ways. We think about time as a one dimensional space. We think about time indicators as points or intervals on this "time line". The activities we perform, the events we participate in, in general all occurrences, have implicitly a temporal and spatial dimension. Occurrences take a certain time to complete, they hold for a certain time interval or they happen at a certain point in time. Also, they happen in a certain region of space, or at a specific location. Because of this, we can use the temporal or spatial dimension of an occurrence to serve as a reference point for another occurrence, just as easily as we use a specific time point or interval, or location in space.

Metonymy allows people to use one well understood aspect of something to stand for the thing as a whole or some aspect of it (Lakoff and Johnson, 1980). We can do this with entities, by substituting

- the object used for the user: *The buses are on strike.*
- the controller for the controlled: *Bush bombed Iraq.*
- the place for the event: *Watergate changed US politics.*

We can also do it with events (Lakoff, 1970). Schank and Abelson (1977) base their text analysis and comprehension using scripts on the hypothesis that people can infer a whole sequence of events from only one mentioned. For example, from the fact that somebody had a meal in a restaurant, we can infer that he entered the restaurant, found a place to sit, ordered food, ate and eventually paid for the meal by giving cash or writing a check or using a credit card.

Metonymy is proof that human communication is based on the assumption that the listener can understand the speaker even if the utterance does not contain all the details. We all have certain experience that helps us fill in some blanks or make associations, or we are just comfortable with a certain manner of speaking, and do not require detailed explanations. This idea is supported also by research in text analysis, who propose that people bring in background knowledge when processing a text, and much of what is missing can be inferred. Dirven and Verspoor (1998) consider that the interpretation of a text is more than the sum of the interpretation of sentences, because of the additional information brought into this process by the reader. In the example:

27 *On our way to the reception, the engine broke down. We were late for the party.*

they claim that people processing this message have no problem understanding that it is the engine of the vehicle used to get to the party that broke, although there is no mention of the vehicle itself. Lyons (1995, p. 266-267) also suggests that people interpret sentences in the context of their background knowledge:

> Usually, we operate with contextual information below the level of consciousness in our interpretation of everyday utterances.

This also helps resolve ambiguities in processed sentences, or it can lead to misinterpretations due to the different expectations of the hearer and the speaker.

Thus, words are more than meets the eye. How much more, and how this extra information is connected to a particular word is a matter of debate.

Nunberg (1978) studied polysemy and the lexicons. Certain word senses cannot be captured in a lexicon, because they arise from the context in which they appear. For example, consider the situation in which somebody shows a key to the parking attendant, and says *This is parked around the corner. key* does not mean *car*, but in that particular situation *key* is used to refer to the *car*. This is a case of "deferred reference", when the speaker refers to a concept B by using the description of another concept A.

This theory evolved into a study of "transfers of meaning" (Nunberg, 1995). According to Nunberg, "transfers of meaning" are "linguistic mechanisms that make it possible to use the same expression to refer to disjoint sorts of things" (p.109). In the particular case of predicate transfer, some property is derived into a new property that is a "noteworthy feature of its bearer".

Nunberg (1978) and Fauconnier (1985) hold the view that metonymy arises from the associations gathered as a result of cultural and existential factors. Cognitive linguists have adopted this idea, and have proposed that metonymy is not merely a figure of speech, but a reflection of the conceptual organization of the world (Lakoff, 1987), (Johnson, 1987).

This associationist idea has been contested, on the basis that the associationist view supports the idea that concept formation is guided by experiential and cultural factors. Linguistic and cognitive development research shows that concept formation seems to rely instead on the built-in human capacity for abstraction (Jackendoff and Aaron, 1991).

Whatever the underlying basis for metonymy, the fact that it is a basic human cognitive process is supported also by psychological research. Gernsbacher (1991) shows that when understanding language, people use metonymy in inferring connections between events.

3.3.3 Language Generation

Work in the field of psycholinguistics is concerned with discovering the processes that humans go through in producing a sentence. In a more technical approach, natural language generation (NLG) is the field of computational linguistics concerned with implementing systems that mimic the human ability of expressing themselves in linguistic form. Although the systems do not claim to replicate the way humans produce utterances, the inspiration comes from psycholinguistic research.

The endeavours in this direction have yielded interesting systems (Hovy, 1996), (Reiter and Dale, 2000). They can be classified according to different characteristics.

- The type of output they produce:
  - **single sentences**. The system concentrates on selecting lexical items which it then combines to produce a sentence.
  - **full texts**. Producing full texts involves a process of text planning, to insure the coherence and informativeness of the text produced. The result of text planning consists usually of discourse structure in a tree-like form. Each leaf contains instructions for a sentence generator to produce a single sentence.
- The technique used for generation:
  - **canned systems**. They produce fixed strings (error messages, warnings, etc.). We find examples of such canned systems in most software tools we use.
  - **template systems**. These systems apply pre-defined templates to produce similar messages. They are used mostly to produce texts with a fairly regular structure, like form letters or business reports. A classic example of such a system is TEXT (McKeown, 1985). It generates a multi-sentence text to answer questions about the structure of a military database. Each type of question has associated a set of possible schemas for the answer (for example, definitions). Based on the knowledge extracted to answer the question, one schema from this set is chosen as a template for the final answer. An extension of this system is TAILOR (Paris, 1993). It uses more schemas and different schema selection criteria.
  - **phrase-based systems**. Such systems use generalized templates. The application of a template resembles the application of a grammatical structure. A matching template, for example [SUBJECT VERB OBJECT] is applied at the

top-most level of the input. Its constituent patterns are recursively expanded into more specific patterns that each match a portion of the input (SUBJECT ↔[DETERMINER ADJECTIVES HEAD-NOUN MODIFIERS]). The process ends when each pattern has been replaced by one or more words. One of the most sophisticated implementations of this technique is MUMBLE (Meteer et al., 1987). Other implementations include RST text structurer (Hovy, 1988) and EES text planner (Moore and William R, 1989).

– **feature-based systems**. Feature-based systems are used to generate single sentences. Each sentence is represented as a unique set of features (POSITIVE or NEGATIVE; QUESTION, IMPERATIVE or STATEMENT; PRESENT, PAST or FUTURE tense, etc.) For each portion of the input the appropriate features are collected. The problems with this type of systems come from the difficulty of maintaining the relationships among features, and performing feature selection. The more features are available, the more complex the input. Some implementations of this sentence generation technique are PENMAN (Mann and Matthiessen, 1985), its descendant KPML (Bateman et al., 1991) and the Functional Unification Grammar framework (FUF) (Elhadad, 1995).

There are shortcomings in each of these text generation systems. They occur at different stages of processing: lexical selection, discourse structure, sentence planning, domain modeling, generation choice criteria. We will expand on those that are relevant to the purposes of this dissertation.

Lexical selection. The input for text generation systems consists of a description of the message in terms of predicates and arguments. The dictionary used by such systems consists of words described in terms of the same predicates. They may also have associated collocational constraints. At the stage of lexical choice, the systems try to find in the dictionary the words that can be substituted for the predicates in the input (Reiter and Dale, 2000). More sophisticated systems will traverse the message several times, rewriting the input structure, and trying to find words that cover as many of the predicates as possible. Search techniques can also be employed to determine the order in which rewriting operations should be performed (for example (Elhadad et al., 1997)).

The lexicons used are still quite limited. When lexicons are extended to comprise different expression possibilities, the lexical selection problem becomes very complex. For an appropriate selection the system must be able to take into consideration different factors about the generated text (the entities, events, etc. already mentioned and referentially available, what is most salient and what stylistic effects the speaker wishes to produce, etc.). For example, consider choosing how to refer to a car: *the car, a car, John's car, the red sport sedan, the car parked under the tree*, etc.

Sentence planning . At the stage of sentence planning, the most critical step to be performed is aggregation. After lexical choice, the system has a collection of *proto-phrases* that it will attempt to put together to form a sentence. There are several aggregation operations:

- **simple conjunction** – combine structures using *and* or contrastive conjunctions.

  There was a mild spell from the 5th to the 9th.
  It was cold at night from the 17th to the 25th.
  ↔ *There was a mild spell from the 5th to the 9th, and it was cold at night from the 17th to the 25th.*

- **conjunction via shared participants** – combine two or more informational elements that share argument position with the same content. The shared content will be realized only once. Shared elements correspond to complete top-level grammatical constituents.

  The month was colder than average.
  The month was relatively dry.
  ↔ *The month was colder than average and relatively dry.*

  January was colder than average.
  February was colder than average.
  ↔ *January and February were colder than average.*

- **conjunction via shared structure** – a situation similar to the one above. The difference is that shared elements do not necessarily correspond to complete top-level grammatical constituents.

  January was drier than average.

January was colder than average.
↔*January was drier and colder than average.*

- **syntactic embedding** (or hypotactic aggregation) – an informational element that might have been realized as a separate major clause is instead realized by means of a constituent subordinated to some other realized element. There are several approaches to syntactic embedding:
  - incorporate an informational element using a nonrestrictive relative clause;

    September only had 30 mm of rain.
    It is usually our wettest month.
    ↔*September, which is usually our wettest month, only had 30 mm of rain.*

  - more sophisticated embedding systems. For example, MAGIC (McKeown et al., 1997) would aggregate the following sentences:

    The patient's last name is Jones.
    The patient is a female.
    The patient has hypertension.
    The patient has diabetes.
    The patient is 80 years old.
    The patient is undergoing CABG.
    Dr. Smith is the patient's doctor.

    ↔*Ms. Jones is an 80-year-old, hypertensive, diabetic, female patient of Dr. Smith undergoing CABG.*

  In order to produce an appropriate syntactic embedding, the system requires representations that encode not only syntactic properties, but also relevant semantic information.

Text generation systems must decide when aggregation operations should be performed, and how to choose among options. There are two competing pressures: conciseness and syntactic simplicity. Aggregation tends to make the messages more compact and concise, but the syntactic structure becomes more complex. On the other hand, a sentence with a simpler structure is easier to understand. It seems to be more acceptable to perform aggregation when the messages being aggregated are semantically related. This source of information remains relatively unused in work on aggregation within NLG. Aggregation rules can be identified by analyzing a corpus for the types of aggregations performed by people in a specific genre of texts (Reiter and Dale, 2000).

There are other sources of sentence aggregation strategies: psychological research on reading comprehension and guidelines and style manuals for human readers. The problem with using psychological findings is that it is often difficult to extract detailed and automatable rules (as opposed to general observations) from this literature.

Scott and de Souza (1990) have hypothesized a number of heuristics, expressed in terms of relations from Rhetorical Structure Theory (Mann and Thompson, 1988), which are at least partially motivated by psychological research.

- **when to aggregate**:
  - aggregate messages linked by a rhetorical relation that is expressed by an explicit cue word.
  - do not express more than one rhetorical relation within a single sentence
- **how to aggregate**: embedded clauses should only be used for Elaboration relations.

Language generation systems start with an encoding of the text to be produced, in terms of the same "primitives" – predicates and arguments – as the words in the lexicon. The expressions that will be produced are limited by the links that can be established between the predicates and their arguments in the text, and the ones in the lexicon. Several word choices are available, depending on how many of the input predicates their definitions cover. Producing and choosing between different possible expressions for the same thing – based on knowledge about previously referred entities, pragmatic considerations, etc – is still a research topic. Another type of producing equivalent expressions for the same underlying definition consists of aggregation of several smaller units into one more complex structure. The application of this phenomenon applies mostly to conjoining phrases. Producing some more complex structures through hypotactic aggregation is restricted for now at combining simple expressions into a larger noun phrase structure.

3.3.4 Grammatical Accounts of Equivalent Syntactic Forms

Syntactic theories also give an account of mapping the same deep structure onto different surface forms.

This started with the study of *transformations*. Harris (1970) was the first to concern himself with this topic in a project on classifying text analyzes according to their informational content. For such an operation to be possible, it was necessary to eliminate textual incoherences and reduce the complexity of the text. *Transformations* were designed to achieve these goals, while preserving the meaning and the grammatical relations (subject, object, etc.).

Harris's transformations operated at the level of the sentence, were based on descriptive rather than empirical facts, and theoretical abstractions and generalizations played a minor role. Here are some examples:

- Transformations which change a sentence of one grammatical form into another grammatical form. For example, active to passive transformations.
  Mary read the book. ↔ *The book was read by Mary.*

- Transformations which change a sentence containing some noun N_1 into a form which is not a sentence, but may be considered grammatically a noun-modifier construct.
  Books are interesting. ↔ *... interesting books ...*

- Transformations which change a sentence into a form which is in general not a sentence, and which enters as the subject or object into another (host) sentence. For example:
  He read the note. ↔ *... his reading of the note ...*

- Transformations which change a sentence into a form which is in general not a sentence, and which is adjoined as a modifier to another sentence. For example:
  The hour is late. ↔ *... the hour being late ...*

One can establish whether two sentences are a transformation of each other by comparing individual co-occurrences of morphemes.

The idea of *transformations* was adopted by Harris's student, Noam Chomsky, who used it to develop a theory of grammar (Chomsky, 1966). The transformations that Chomsky proposed operated on structures rather than sentences. According to this theory, two sentences are related if they are derived from the same *deep structure*.

Deep structures correspond to an innate capacity of the human brain, and were considered to constitute a universal grammar. On top of this deep structure there are rules – transformations – that act upon the deep structure to produce sentences. The end result of a transformational-generative grammar is a surface structure that, after the addition of words and pronunciations, is an actual sentence of the language. All languages have the same deep structure, but they differ in surface structure because of the application of different rules for transformations, pronunciation and word insertion.

Transformations are defined in terms of movements of constituents in a tree-like structure.

While attempting to better specify transformations and situations in which they may apply, this work has evolved into another theory of grammar, called Government and Binding (Chomsky, 1982). The movement of constituents inside the structure is restricted by the head of the phrase that governs the constituent.

Further analysis of the theory of generative grammar has lead to yet another theory – the minimalist program (Chomsky, 1995), (Radford, 1997). What differs from one theory to the next are the definitions of possible transformations, and the constraints on their applications.

These transformations do indeed explain the existence of various sentences to express the same deep meaning, but the syntactic level on which the constituents of these sentences are represented does not change. Moving entire constituents inside a parse tree does not change the nature and form of these structures. We are interested in phenomena that explain how the same concept surfaces in expression that pertain to different syntactic levels.

The only exception from the observation above is nominalization. Although nominalization is a transformation accepted by Harris (1970), in transformational grammars it is not. It is considered one cannot produce a derived nominal and its verbal counterpart from the same deep structure (Chomsky, 1970). Another degree of nominalization, gerundive nominals, can.

28 *John is eager to please.*

29 *John's eagerness to please* (derived nominal)

30 *John's being eager to please* (gerundive nominal)

Derived nominals are closer to noun phrases than sentences, judging by their behaviour relative to modifiers (can they take determiners, adjectives, etc.) and pluralizing. Structurally they are close to sentences, however.

Chomsky proposes that the deep structure capture the difference between derived nominals and sentences, and the lexicon carry the similarities by assigning one lexical entry to alternative forms (for example easy and easiness, eager and eagerness, etc.). This is what has been called the lexicalist analysis of derived nominals.

On the other hand, there are those who support a strong structural parallel between a nominal phrase and the corresponding clause:

> The claim is however that we can match elements of the noun phrase (head, modifiers, determiners) with elements of clause structure, considered semantically in terms of the verb and its associated participant roles of agentive, affected, etc. (Quirk et al., 1985)

The conclusion of this literature overview is that there are changes that can be applied to a sentence that allow it to preserve its meaning, for example transformations (in the meaning assigned by (Harris, 1970) and (Chomsky, 1966)). The view that different syntactic expressions convey the same meaning is contested by researchers who propose that different syntactic forms convey (slightly) different meanings, and they have evolved because the existing forms did not meet all the communication goals of the speaker.

It is true that elements of the meaning of an utterance, such as focus or implications, change from one expression to the other. However, the underlying concepts and the links between them may remain the same. This is the level we are interested in, and relative to which we propose that different surface forms of ideas can be the same. We will not concern ourselves with phenomena like the transformations proposed by the supporters of the transformational grammar paradigm. These transformations do indeed change the surface form of the sentence, but not with respect to the syntactic level on which its constituents are represented. Moving entire constituents inside a parse tree does not change the nature and form of these constituents. We are interested in phenomena that account for the same concept surfacing in language in expressions that belong to different syntactic levels.

3.3.5 Lexicology

We have shown until now that various fields of research (language production, natural language generation, generative grammar) are interested in the way in which ideas surface in language, and the different forms they can take. There is one other level which can account for the same, or a similar, concept surfacing in different forms. It is the lexical, or word-formation, level.

The distinction between the grammatical and the lexical level is language specific.

> What is encoded lexically (lexicalized) in one language may be encoded grammatically (grammaticalized) in another (Lyons, 1995).

Languages with free word order tend to be more lexicalized than languages like English which have fixed order. Some of the distinctions made in English by using a specific ordering of constituents are achieved in other languages by inflection.

As in the other work reviewed, we are interested in different surface forms that express the same, or very similar, concepts.

One obvious possibility is given by synonyms. The notion of synonymy is a much debated issue. In the literature distinctions are made among near, partial and absolute synonymy (Lyons, 1995). The differences occur when the senses of two possibly synonymous words or expressions cannot all be mapped onto each other, or they are not substitutable one for the other in all possible contexts. Lexical synonymy is not relevant to our study. Replacing a word with another under the same part of speech has no impact on the structure of the sentence.

Other possibilities pertain to the domain of word-formation. A complex expression can be compacted in one word, as in the examples: *the one who teaches* ↔*teacher*, *has no colour* ↔*colourless*, etc. Phenomena of word formation show how a complex idea can be conveyed through one word. This is relevant to our discussion on different surface forms for the same concept.

3.4 Surface Forms

It is not our purpose to propose models or rules of language production, or a generative grammar for surface expressions. The review serves to support the view that the same idea can surface in different syntactic forms. We now look for mechanisms or processes that account for the different realizations. We work with different texts, and look at them from the point of view of semantic relations. We look for reasons why we can find the same semantic relations on different syntactic levels. Each semantic relation imposes its own restrictions on the concepts whose interaction it describes. We identify these restrictions and search for surface expressions of the same concepts, or the same type of concepts, with respect to the semantic relations in which they are involved. We then analyze the different surface forms, and try to find a systematic account of the differences.

Semantic relations have been analyzed by the syntactic level to which they belong. It is a valid distinction; on different levels, semantic relations describe the link between different types of syntactic structures:

- At the **clause level** the link is between simple clauses.
- At the **intra-clause level** the link is between a verb and its arguments.
- At the **noun-phrase level** the link is between a noun and its modifiers.

Because in this dissertation we propose a view that does not distinguish relations by the syntactic level, we want to break this barrier. Semantic relations connect occurrences, an occurrence and the entities involved, an occurrence or an entity and its attributes. For example, an Agent relation describes the link between an occurrence and the one who is performing it. An Effect relation describes the link between two occurrences, one of which is the effect of the other.

We will show that an occurrence can be expressed through a verb, a nominal (gerund, deverbal noun, true noun) or a deverbal adjective. It is true that words in different parts of speech, by their nature, stand for different types of concepts. For example nouns express entities, or rather, what we perceive as entities we usually express through nouns. Occurrences are expressed through verbs, or clauses. However, when two connected concepts surface in various forms, the relation we perceive between them remains the same, even if the syntactic form imposes a holistic view, or a reification, of an occurrence, or an attributive view associated with an entity, as it would be the case in the example:

31 *parental control ↔The parents control ...*

When discussing semantic relations, talking about syntactic levels can be misleading. There are semantic relations that connect two clauses, but appear at the intra-clause level rather than at the clause level. This is the situation for verbs like *know* or *say*, which can take a clause as a syntactic object:

32 *I know what you did last summer.*

The semantic relation that connects *know* and the clause *what you did last summer* is a Case, not a clause-level relation. Yet the connection is between two occurrences, where one is an element (for this particular example the *Object*) of the other.

Even if syntactic levels are not central to this reasoning, we must somehow anchor the semantic relations in syntax, to be able to analyze the different syntactic expressions that concepts can take. We will do that by reorganizing the type of syntactic units linked by semantic relations.

We reorganize the type of syntactic structures we deal with in a sentence, to mirror the pairing of units for semantic relations assignment. We will consider the following types of syntactic structures, built around open-class words. These structures consist of a center which is an open-class word W, and all the syntactic units in the sentence with which it is connected through semantic relations. These units are syntactically subordinate to W.

- **Verb-centered**. We refer here to clauses, as opposed to verb phrases. The reason for choosing a clause is to include the subject of the clause which is not a part of the verb phrase, but which is connected to the verb through a semantic relation, such as, for example, AGENT.

- **Nominal-centered**. This type of structure covers noun-phrases. The center is a nominal (gerund, deverbal or true noun), and the connected units are the nominal's modifiers in the sentence.

- **Adjective-centered**. Although there exist complex adjective phrases, for the present analysis this structure will consist of the adjective alone. The reason for this restriction is that there are no semantic relations in which an adjective is not connected to a noun or verb. In the remainder of this chapter we will refer to this type of structure as simply *adjective*.

- **Adverb-centered**. For similar reasons, this structure will consist of the adverb alone. It will appear in relations with nouns or verbs. We will refer to this structure as *adverb*.

From the closed-class words, pronouns can appear as elements in a semantic relation. They constitute a special case with respect to surfacing phenomena. We will discuss them in Section 3.4.3.4.5.

We will show first how the view of semantic relations between syntactic structures pertaining to different syntactic levels can be changed into a view of relations between syntactic units centered on words in different parts of speech. Some of the changes are actually trivial: a clause is a verb-centered structure, and a relation between two clauses will be mapped onto a relation between two verb-centered structures.

We will show afterward how the problem of the identity of a semantic relation on different syntactic levels is reduced to the idea that concepts, or types of concepts, can surface as verb-, noun-, adjective- or adverb-centered structures.

3.4.1 Semantic Relations Between Structures Centered on Open-Class Words

A noun phrase is a nominal-centered structure. The relations we assign inside a noun phrase are between the head noun and its modifiers. The modifiers can be adjectives (*happy girl, someone happy*), adverbs (*far-away place, the people behind*), participles (*hidden treasure, interesting book*), other noun phrases (*examinations board*) or even clauses (relative (*the man we saw yesterday*), non-finite (*the car standing outside the station*), appositive (*the saying that absence makes the heart grow fonder*) or premodifying (*[he asked] we don't know how many people*)). If the modifier is a prepositional phrase, a semantic relation is assigned between the head noun and the prepositional complement. The preposition is used as a marker (or indicator) for the semantic relation.

In conclusion, the modifiers of a noun phrase can be any of the structures we deal with: a verb-, noun-, adjective- or adverb-centered structure.

A clause consists of a verb and its arguments. A verb's arguments may be noun phrases (*the woman sang*), prepositional phrases (*[the actor] walked onto the stage*), adverbs (*[she] walked slowly*), adjectives (*[the dog] is brown*) and also clauses (*[He] knows where you hid the presents*).

Just as for modifiers in a noun phrase, a verb's arguments can be any open-class word-centered structure (OWCS).

In the light of these considerations, let us take a look at how syntactic structures change when we want to map the same semantic relation onto pairs of units at different syntactic levels.

3.4.1.1 Clause Level ↔Intra-Clause Level, or Clause-Clause ↔Verb-Argument

We want to see if it is possible for a relation between two clauses to be the same as a relation between a verb and its argument.

33 (CL) *The painter listens to music while he mixes his colours.*

34 ↔(IC) *The painter listens to music while mixing his colours.*

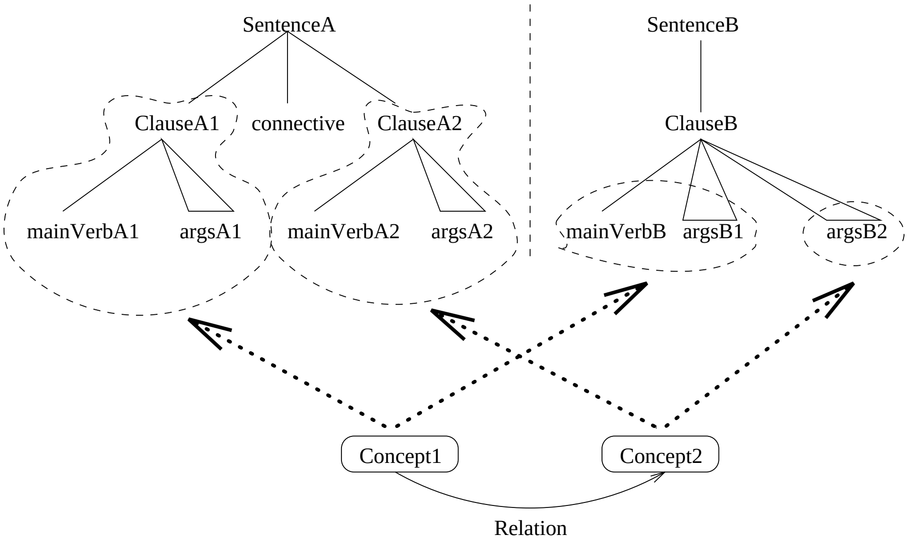


Figure 3.1: Clause-clause ↔verb-argument

A clause is a verb-centered structure, and a verb's arguments can be any type of OWCS.

The semantic relation at the clause level holds between two clauses. From the point of view of OWCSs, the relation holds between two verb-centered structures.

At the intra-clause level, the semantic relation will hold between the main verb and its arguments, or equivalently, between a verb-centered structure and any OWCS.

In order for the semantic relation between two clauses to be the same as the relation between a verb and its arguments, one must find the same semantic relation connecting two verb-centered structures, and connecting a verb-centered structure and an OWCS.

Figure 3.1 shows the projection of two connected concepts at the clause and the intra-clause level. *Concept1* is mapped onto two verb-centered structures, *Concept2* is mapped onto a verb-centered structure and another OWCS.

3.4.1.2 Intra-Clause Level ↔Noun Phrase Level or Verb-Argument ↔Noun-Modifier

Following the same considerations as in the preceding section, we want to see how a relation at the intra-clause level can have instances at the noun phrase level, as in the examples:

35 (IC) *The band practices during lunch hour.*

36 ↔(NP) *lunch-hour practice*

From the point of view of structures centered on open-class words, the situation when the same relation is encountered at the intra-clause and the noun-phrase level can be described as follows: the relation between a verb-centered structure (V) and an OWCS (Arg) can also be found between a nominal-centered structure (N) and another OWCS (Mod). V and N stand not only for the main verb and the head-nominal respectively, but for the main verb and zero or more of its arguments, and the head-nominal and zero or more of its modifiers We analyze all possible mappings of a pair of concepts onto the pairs (V,Arg) and (N,Mod).

1. $V \longleftrightarrow N$, $Arg \longleftrightarrow Mod$. Map the same concept onto the heads of the structures on the two syntactic levels, map the related concept onto modifiers/arguments. This situation is illustrated in Figure 3.2, and examples (37) and (38) show a concrete example of this mapping onto structures pertaining to different syntactic levels.

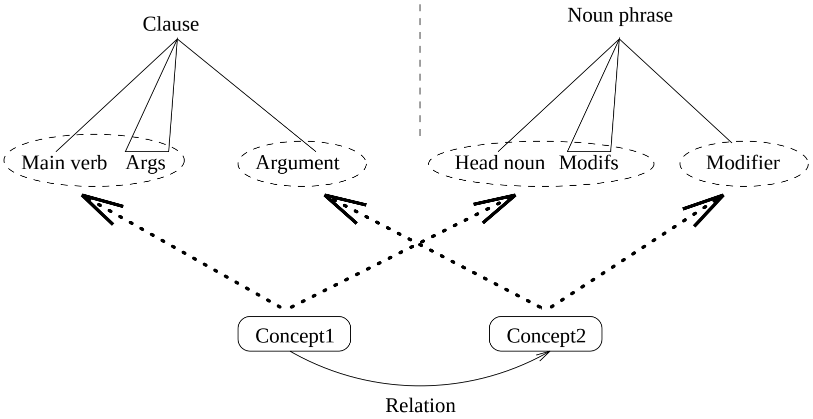


Figure 3.2: Verb-argument ↔noun-modifier: head-to-head mapping

37 (IC) *The parents refused [to sign the consent form.]*

38 (NP) *parental refusal*

In this situation, the verb-centered structure will be mapped onto a nominal-centered structure (mapping the verb onto the head noun), and an OWCS will be mapped onto another OWCS (mapping the argument of the verb onto the modifier of the noun).

2. $V \longleftrightarrow Mod$, $Arg \longleftrightarrow N$. Map one concept onto the main verb and one of the noun's modifiers, map the other concept onto the noun and one of the verb's arguments. Figure 3.3 gives a schematic representation of this mapping.

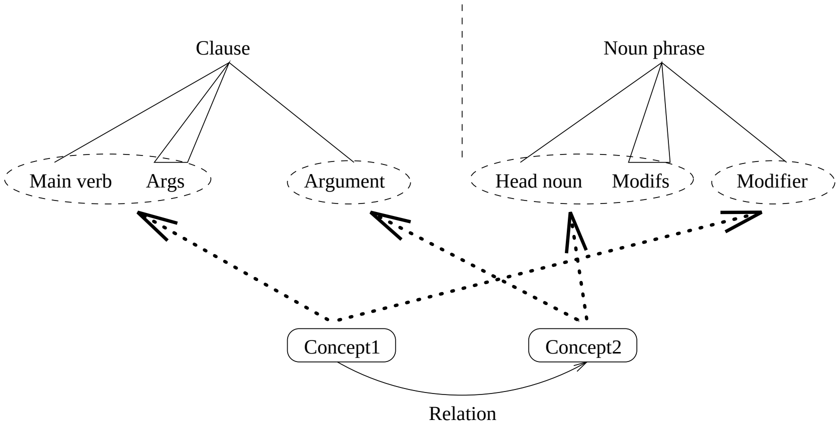


Figure 3.3: Verb-argument ↔noun-modifier: head-to-modifier mapping

For example, consider the pairs of sentences (39,40) and (41,42).

39 (IC) *[They] repaired the engine.*

40 ↔(NP) *repaired engine*

41 (IC) *The liquid has no colour.*

42 ↔(NP) *colourless liquid*

The verb will surface as a modifier, which means that a verb-centered structure will be mapped onto any OWCS. The argument of the verb will surface as the head noun, therefore an OWCS will be mapped onto a nominal-centered structure. Each of these cases is discussed in this section.

3. $V \longleftrightarrow \epsilon$, $Arg1 \longleftrightarrow N$, $Arg2 \longleftrightarrow Mod$. Map one concept onto one of the verb's arguments and onto the head noun, map the other concept onto another argument and one of the noun's modifiers. The main verb serves as a link between its arguments, it does not have a counterpart in the noun-phrase, as shown in Figure 3.4.

For example, since *production* is implicit in a factory, the sentence:

43 (IC) *The factory produces cars.*

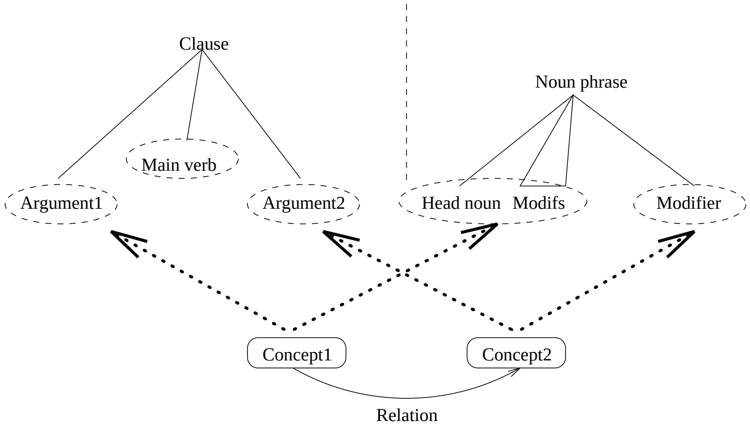


Figure 3.4: Verb-argument ↔noun-modifier: verb deletion

can be compacted into the noun-phrase:

44 ↔(NP) *car factory*

A verb's argument can be any type of OWCS, and the same is true for a noun's modifier. The structure corresponding to one of the verb's arguments will be mapped onto a nominal-centered structure (corresponding to the head nominal), and the other onto the structure corresponding to the modifier of the head nominal.

4. $N \longleftrightarrow \epsilon$, $Mod1 \longleftrightarrow V$, $Mod2 \longleftrightarrow Arg$. As opposed to the case where the verb can be excluded because it serves only as a link between its arguments, the noun is never conceptually empty, or retrievable from the relation between its modifiers. This mapping is not possible.

3.4.1.3 Clause Level ↔Noun Phrase Level or Clause-Clause ↔Noun-Modifier

For a relation between two clauses to surface as a relation between two components of a noun phrase, each of the two clauses must find an expression as elements of a noun phrase.

There are two situations here, depending on the form of the noun phrase:

1. The noun phrase is a paratactic construction (its constituents are at the same level in the structure). Figure 3.5 illustrates this situation.

   Sentences (45) and (46) show a concrete example of this mapping,

   45 (CL) *If you eat sweets before you eat dinner [you spoil your appetite.]*

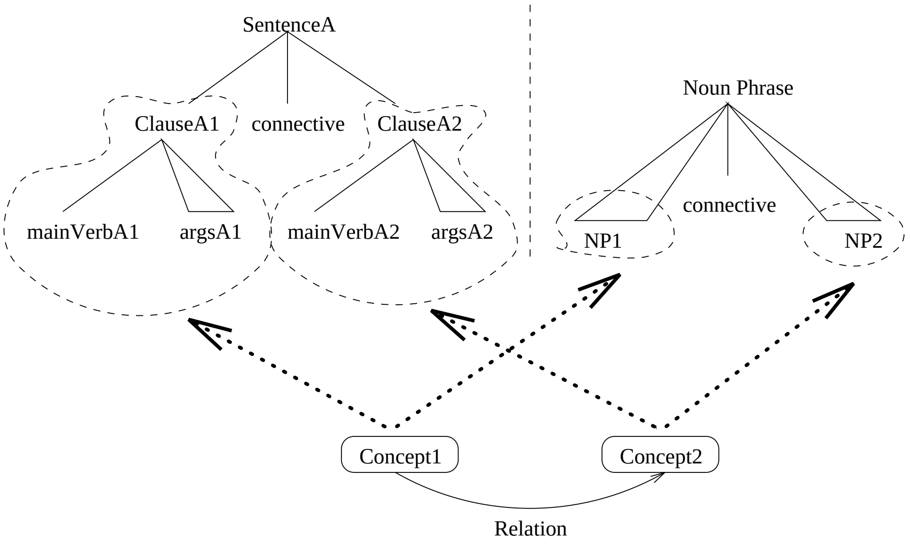


Figure 3.5: Clause-clause ↔noun-modifier: paratactic construction

46 ↔(NP) *sweets before dinner [spoil your appetite]*

2. The noun phrase is a hypotactic construction (one constituent is subordinate to the other), as shown in Figure 3.6.

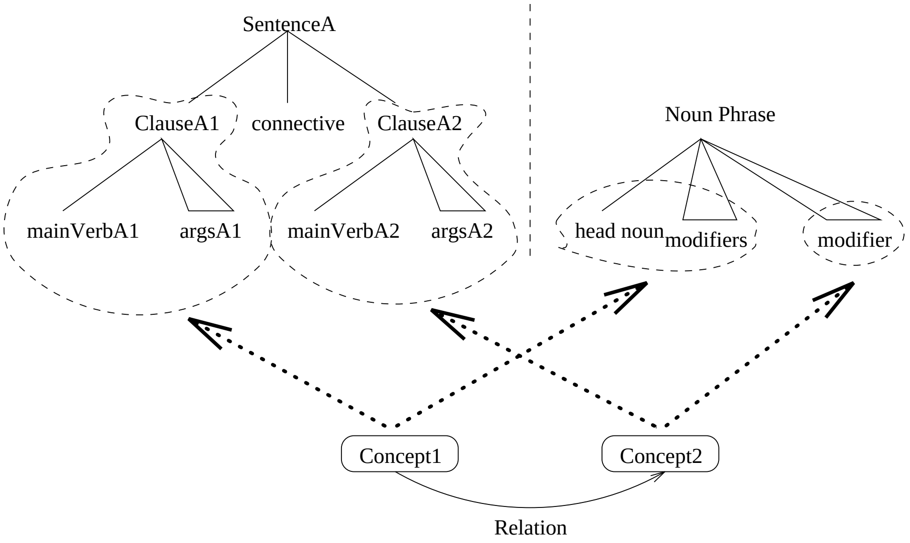


Figure 3.6: Clause-clause ↔noun-modifier: hypotactic construction

This mapping is illustrated in examples (47) and (48).

47 (CL) *The student was anxious because he was writing an exam.*

48 ↔(NP) *exam anxiety*

A clause is a verb-centered structure, a noun is a nominal-centered structure and its arguments can be any type of OWCS. In order to have an instance of the same relation at the clause and noun-phrase level, the same concept should have a verb-centered and a nominal-centered realization, and the concept with which it is connected should be expressed as a verb-centered structure and as another OWCS.

3.4.2 Mapping a Concept Onto Different Syntactic Structures

The analysis in Section 3.4.1 shows that in order to show that the same semantic relation can have instances at different syntactic levels, it must be shown that a concept can surface as two different open-class word-centered structures (OWCSs).

We will take each mapping between OWCSs, and look for phenomena in language that allow a concept to take such distinct syntactic forms.

Table 3.1 shows all possible mappings that arise from this discussion, as all combinations of two OWCSs. Since we will talk about mapping the same concept onto different syntactic structures, the order of the structures in the discussion does not matter.

Table 3.1: Possible structural mappings

Mappings of a concept onto pairs of structures	
(verb-centered,	nominal-centered)
(verb-centered,	adjective)
(verb-centered,	adverb)
(nominal-centered,	adjective)
(nominal-centered,	adverb)
(adjective,	adverb)

One of these mappings is not possible. The same concept cannot be mapped onto a verb-centered structure and an adverb-centered structure. An adverb serves to clarify some aspect of the occurrence described, but it cannot stand for a whole occurrence.

In order for an adverb to conjure an entire occurrence in the mind of the reader, it must be strongly associated with the occurrence lexically or conceptually. In order to have a strong lexical association between an adverb and a verb, some lexical or grammatical phenomena must relate the two words. But there is no word-formation process that accounts for the formation of an adverb directly from a verb phrase (Quirk et al., 1985). Conceptually, adverbs are too general to stand for a particular occurrence.

In the sections that follow we will analyze concept mapping onto every possible pair of structures.

3.4.2.1 Mapping a Concept Onto a Verb-Centered Structure and Onto a Nominal-Centered Structure

Since we study different expressions of the same, or very similar, concept from the point of view of its semantic relation with another concept, we will not restrict the examples presented to the concept that is mapped onto syntactic structures. We will present it instead with the whole context that is, with another concept, relative to which the surface expressions convey very similar meanings.

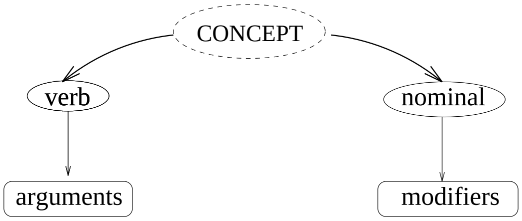


Figure 3.7: Mapping a concept onto a verb-centered structure and onto a nominal-centered structure

Mapping a concept onto a verb-centered structure and onto a nominal-centered structure (Figure 3.7) can take different forms.

We have defined the concept as covering occurrences, entities and their attributes. Occurrences in turn have been defined as covering the entire situation they describe, including participants and attributes. Entities also include their attributes. From this perspective, concepts are not atomic. In the discussion that follows we need to look inside the concept and see how its various parts are mapped onto syntactic structures. We will distinguish *the core*, the possible *participants* and *attributes* as parts of a concept. The *core* will correspond to the concept stripped of all information regarding participants and attributes.

- Map the core of the concept onto the heads of the two structures, and its participants and qualifying attributes onto the arguments or modifiers respectively (Figure 3.8).

  The concept that is produced as a verb can be produced as a nominal:

  49 *The painter listens to music while* **he mixes his colours**.

  50 ↔*The painter listens to music while* **mixing his colours**.

  The two linguistic phenomena that account for this possibility are **nominalization** of the verb and **verbalization** of the noun.

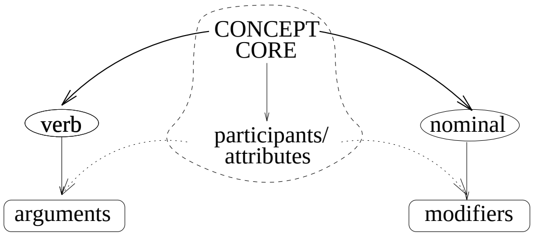


Figure 3.8: Mapping a concept core onto the heads of the structures

- Map the concept onto the verb and onto one of the nominal's arguments, map the related concept onto the nominal and onto one of the verb's arguments (Figure 3.9).

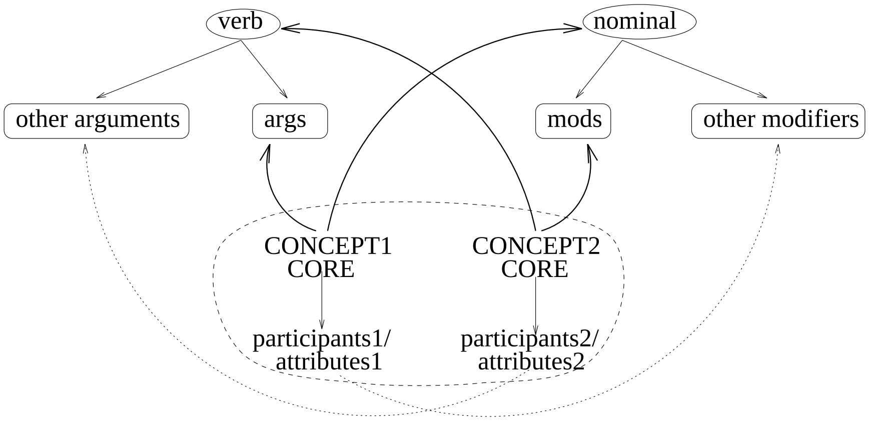


Figure 3.9: Mapping a concept core onto a head and an argument

The event expressed by the verb can be equivalently expressed as a noun's argument:

51 *The ship sank.↔sunken ship*

In example 51 the relation between the two occurrences of *sinking* corresponds to a relation between a verb and a deverbal adjective. Generally, we have the mapping of a concept onto a verb and an argument of a noun-phrase, which is an OWCS. All possible mappings are discussed in this section. The same is true for mapping the same concept onto a noun and a verb's argument.

- The concept expressed by a combination of the verb and some of its arguments can surface as a noun (Figure 3.10).

  This situation can be explained by the phenomena of **word formation**. For example:

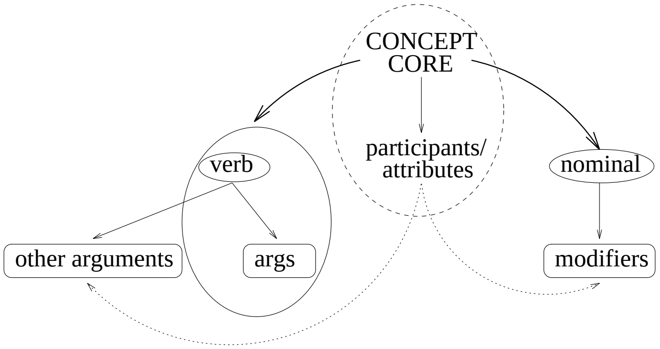


Figure 3.10: Mapping a concept onto a verb and some of its arguments and onto a nominal

the one who builds (vb+subj) ↔*builder* (noun)

the one who was appointed (vb+obj) ↔*appointee* (noun)

- The verb has no equivalent expression in the noun phrase (Figure 3.11).

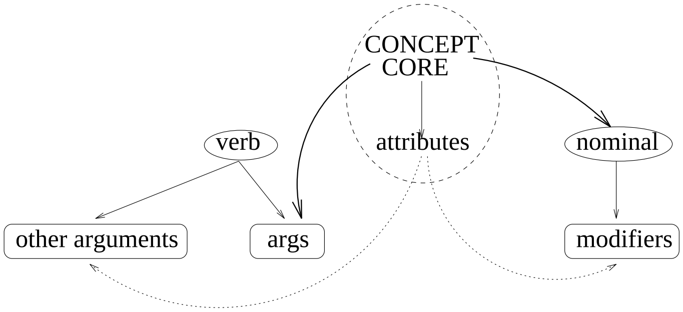


Figure 3.11: Mapping a concept onto verb arguments and onto a nominal and modifiers

There are three explanations for this situation:

1. **Deletion**. The verb is deleted. A combination of its arguments is enough to convey a similar meaning:

   52 *The house <u>was built with bricks</u>. ↔<u>brick</u> house*

2. **Metonymy**. The whole occurrence expressed by the verb-centered structure was replaced with one of its arguments:

   53 *The student was anxious because <u>he was writing an exam</u>.*

   54 *The student was anxious because of <u>the exam</u>.*

3. **Equivalence**. In the given context, the concept expressed through a verb-centered structure can be replaced by another concept expressed through a nominal-centered structure, without changing the nature of the relation with the context:

   55 *We will have tea when you arrive.*

   56 *We will have tea at 5 o'clock.*

   In this case, the relation that is maintained despite the change of concept is a temporal relation. The type of the concept is the same in the context of the sentence. Both *5 o'clock* and *you arrive* identify a point in time.

In the discussion of mapping concepts onto various syntactic forms we have not talked much about the arguments/modifiers in a syntactic structure. There is a difference between the way verbs and nouns behave with respect to their arguments. Verbs subcategorize for certain arguments, while for nouns they are all optional (Quirk et al., 1985). Therefore, some of the verb's arguments may surface as noun's modifiers, but there is not necessarily a one-to-one mapping. The arguments give more information, clarify the circumstances and the manner of an occurrence, etc. The arguments in each structure that are important for the mapping were shown in the mapping process. The ones that are not shown can be mapped according to the processes described in Section 3.4.2. The arguments in each structure (verb- or nominal-centered) are themselves noun-, adjective-, adverb- or verb-centered structures. We analyze all possible mappings of a concept onto pairs of structures. These mappings also apply to concept attributes and participants that surface as the arguments in a verb-centered or nominal-centered structure.

3.4.2.2 Mapping a Concept Onto a Verb-Centered Structure and Onto an Adjective

There are two possibilities of mapping a similar concept onto a verb-centered structure and onto an adjective.

- Map a concept onto the verb and onto the adjective (Figure 3.12).

  57 *The ship sank.* ↔*sunken ship*

- Map a verb and one (or more) of its arguments onto an adjective(Figure 3.13).

  58 *The liquid has no colour.*↔*colourless liquid*

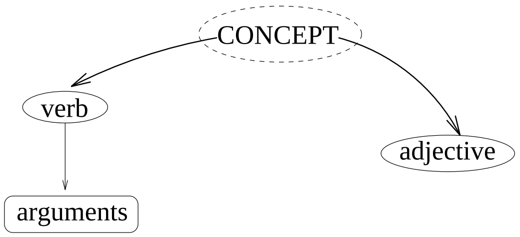


Figure 3.12: Mapping a concept onto a verb and onto an adjective

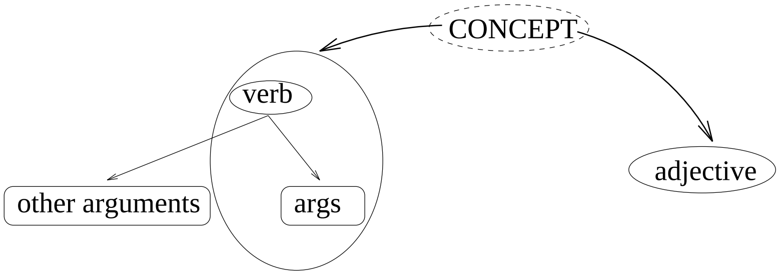


Figure 3.13: Mapping a concept onto a verb and some of its arguments and onto an adjective

The phenomena that explain the mapping of a concept onto a verb-centered structure and onto an adjective pertain to **word formation**, in particular, formation of adjectives from verbal expressions, and formation of verbs from adjectival expressions.

3.4.2.3 Mapping a Concept Onto a Nominal-Centered Structure and Onto an Adjective

- Map the same concept onto the head noun and onto the adjective (Figure 3.14).

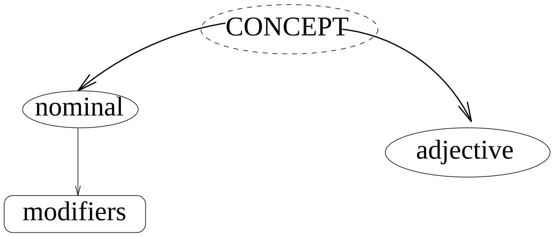


Figure 3.14: Mapping a concept onto a nominal and onto an adjective

59 *The parents refused ...* ↔*parental refusal*

60 *... writing with style ...* ↔*stylish writing*

- Map the same concept onto the head noun and one or more of its arguments and onto the adjective (Figure 3.15).

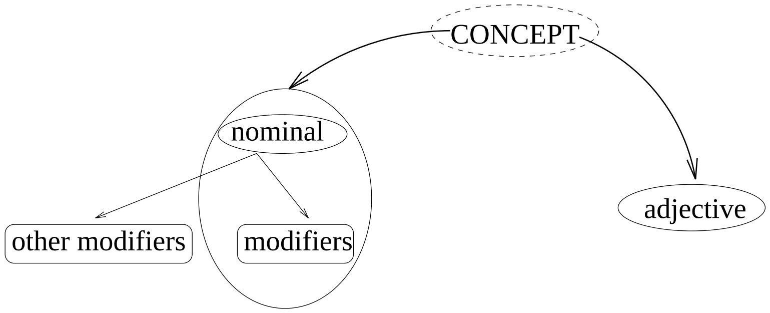


Figure 3.15: Mapping a concept onto a noun and some of its arguments and onto an adjective

This theoretical possibility seems to have no realization in English.

Noun and **adjective formation** phenomena (adjective nominalization, adjectivalization of the noun, derivation) account for the mapping of the same concept onto noun- and adjective-centered structures.

3.4.2.4 Mapping a Concept Onto a Nominal-Centered Structure and Onto an Adverb

In a context that expresses direction, we find adverb formation phenomena that allow a nominal expression to surface as an adverb. For example, we can equivalently say *towards home* or *homeward*, *towards north* or *northward*, etc. In all these cases, the noun must be a prepositional complement, and the entire prepositional phrase has an alternative adverbial expression (Figure 3.16).

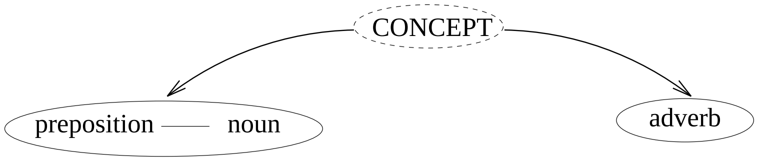


Figure 3.16: Mapping a concept onto a prepositional phrase and onto an adverb

A quantified noun phrase (*every day*) can also be mapped onto an adverb (*daily*). Figure 3.17 shows this situation.

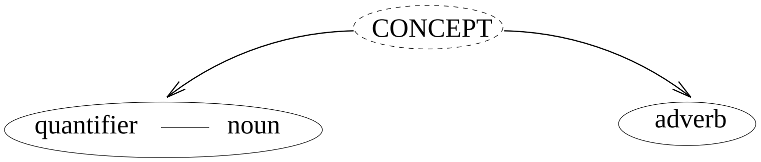


Figure 3.17: Mapping a concept onto a noun-phrase and onto an adverb

3.4.2.5 Mapping a Concept Onto an Adjective and Onto an Adverb

Figure 3.18 shows the situation when the same concept is mapped onto an adjective and onto an adverb.

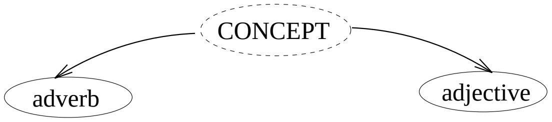


Figure 3.18: Mapping a concept onto an adjective and an adverb

For example:

61 *He was driving weekly to the mountains.*

62 ↔*weekly drive*

3.4.3 Surfacing Phenomena

Based on the analysis of syntactic structures onto which concepts can be mapped, we have identified phenomena that account for the existence of different surface forms of concepts. In this section we will take each phenomenon, and analyze it from the point of view of semantic relations. We will show that the existence of these phenomena supports the idea that semantic relations can be the same on different syntactic levels.

3.4.3.1 Metonymy

That words are polysemous is a fact captured in all dictionaries and lexical resources concerned with word meaning. However, not all senses of a word can be captured in such a manner. Sometimes words acquire a very specific meaning based on the context in which they are used. In such a context, people can infer that a specific word refers to something that is not there. This issue has been a topic of research starting with (Nunberg, 1978), and has been a matter of debate in various circles ever since.

The phenomenon by which something is referred to through a more common or well known aspect of itself is termed *metonymy*. It is a pervasive phenomenon in language.

We usually have no problem understanding the following sentences, even though some of them may have several possible readings:

63 *John started the experiment this morning.*

64 *Sweets before dinner spoil you appetite.*

65 (a person showing the parking attendant his car key) *This is parked around the corner.*

Metonymy is defined as a linguistic device by which a concept is referred to by one of its attributes or something that is associated with it:

> **metonymy** [...] the substitution of a word referring to an attribute for the thing that is meant, as for example the use of *the crown* to refer to *a monarch.* (Hanks, 1986)

While metonymy allows for the substitution of an attribute for the thing or the other way around, synecdoche allows for the substitution of the part for the whole, and the whole for the part:

> **synecdoche** [...] a figure of speech in which a part is substituted for a whole, or the whole for a part, as in *50 head of cattle* for *50 cows*, or *army* for *a soldier.* (Hanks, 1986)

The term **synecdoche** describes more appropriately the phenomena we analyze. However, it is not much used in the literature, and the definition of metonymy was stretched to include synecdoche as a special case (Dirven and Verspoor, 1998), (Lakoff and Johnson, 1980). We will use the term **metonymy** as well.

Metonymy accounts for different surface realizations of occurrences. An occurrence can be realized through one of its participants. Here is an example: we define the Effect relation as a relation between two occurrences, the *Cause* and the *Effect.* The proposition expressing the *Cause* is true, and this makes the *Effect* proposition true.

66 (CL) *The student was anxious because he was writing an exam.*

This is an example of the Effect relation at the clause level. The *Cause* proposition is *he was writing an exam*, the *Effect* is *the student was anxious.* The *Cause* proposition is true, and, as a consequence, the *Effect* proposition becomes true. Syntactically speaking, both *Cause* and *Effect* are expressed by occurrences – action for *Cause*, state for *Effect* (in this example (66)).

It is possible that the speaker does not know exactly what about the exam makes the student anxious – writing it, answering it orally, thinking about it, or some other event. Then the speaker might choose the following expression:

67 (IC) *The student was anxious because of the exam.*

In this situation the relation is still EFFECT. The *Effect* proposition is still expressed by a state occurrence, but the *Cause* is now just an entity – *exam*.

Previously we have established that causal relations hold between two occurrences. How does this example fit the relation's requirements? There is a coercion from *exam* to the occurrence of WRITING/ANSWERING AN EXAM that explains why the *Cause* is an occurrence that is not fully expressed.

We regard the occurrence as a whole whose parts are its constituents – the main verb, the participants (*Agent, Object*, etc.), the qualifiers, etc. In light of this interpretation, we can extend the definition of metonymy to include the substitution of a part of an occurrence for the whole.

The phenomenon that explains why example (67) presents a causal relation that links two occurrences is **metonymy**: a part (in this case *exam*, which may be the *Object* of the fully expanded occurrence) stands for the whole.

We can design experiments to retrieve some of the missing information from corpora, through collocation information. In the example:

68 *Coffee after 5 o'clock doesn't let you sleep.*

the reader can infer that we mean:

69 *<u>Drinking coffee</u> after 5 o'clock doesn't let you sleep.*

In certain cases such information can be retrieved or inferred with a certain probability from corpus analysis. We have found collocation information associated with the noun *coffee* extracted from the British National Corpus as part of the WASPS project (Kilgarriff and Tugwell, 2001). From all the verbs it appears with as a syntactic object, the most frequent one is *drink*.

There are however situations in which corpora analysis does not help, because the event associated with a certain word is not an usual event, but something that arises from the context of the sentence. For example, consider the sentence:

70 *Billy began the book.*

This sentence by itself conveys the same meaning as the following, more extensive sentence:

71 *Billy started to read/write the book.*

If sentence (70) follows a text fragment in which it is established that Billy is in fact a goat, then we would rather be inclined to consider sentence (70) as equivalent to sentence (72).

72 *Billy began eating the book.*

Here is an alternative explanation of the fact that in a certain context a word conjures up a whole occurrence in the mind of the reader: this extra information is so closely linked to the meaning of a word that it should be part of the lexicon. This is the approach proposed by Pustejovsky in his generative lexicon (Pustejovsky, 1995). Pustejovsky proposes that information about the meaning of a word in context should be available through a lexicon that comprises explanations about the concept x behind each word. The explanations cover four aspects:

- **Constitutive**. What x is made of.
- **Formal**. What x is.
- **Telic**. What x's function is.
- **Agentive**. How x came into being.

These four aspects form the *qualia* structure of a lexical item. The *qualia* "constitute the necessary modes of explanation for understanding a word or phrase" (Pustejovsky, 1998). In order to reach the meaning of a word in context, one makes use of one or more of the aspects associated with it through the lexicon.

For the noun *coffee* in example (68), the *telic* role would bring the information that *coffee* is for *drinking.*

The generative lexicon approach could not explain the situation presented in examples (70) and (72). All information that is incorporated in a lexicon is generic, and it cannot address pragmatic considerations that arise from the specific context in which the word is used.

Metonymy covers a broader array of phenomena. However, in replacing the part by the whole or vice versa, or an attribute for the entity and vice versa, we replace a noun phrase with another noun phrase which does not change the syntactic structure analyzed. The level on which the semantic relation was expressed is the same. Here we are interested only in phenomena which show semantic relations on different syntactic levels.

3.4.3.2 Equivalence

We have seen that metaphor allows us to conceptualize one thing in terms of another, through perceived similarities. We conceptualize time as a one-dimensional space, onto which we can represent occurrences and specific time points or intervals. Occurrences have an implicit temporal and spatial frame: they unfold during a specific time interval or at a certain point in time, and they unfold in a certain location in space. The temporal dimension of occurrences can be taken as a reference for other occurrences just as well as an explicit time indicator can. The same is true for spatial expressions, which can be an explicit location, or a location identified though an occurrence that unfolds in that particular space.

The manner in which an occurrence unfolds can also be defined in terms of another occurrence, or through an adverbial expression.

We will name the phenomenon that allows for the same type of concept (temporal, spatial or manner expressions) to surface in language in various forms **equivalence**.

In the following example:

73 (CL) *They practice while others have lunch.*

we fix the time when the *practice* occurrence takes place, by expressing another occurrence that unfolds during the same time interval. Or we can choose to use a specific time interval instead:

74 (IC) *They practice during lunch hour.*

In this case the time interval during which the *practice* occurrence is unfolding is a definite time interval: the *lunch hour.*

In discussing temporal aspects of sentences, among other things, Quirk et al. (1985) look at the use of *till* and *until* in realizing adjuncts of forward span. Their observation is that *until* usually introduces a clause, whereas *till* introduces a prepositional phrase. This is relevant because it shows that forward span has alternative expressions as a multi-clause or as a simple sentence:

75 (CL) *They will live in Chicago until William finishes his thesis.*

76 (IC) *She will be working till nine o'clock.*

It is not important for our work that sometimes *till* can be used to introduce a clause, just the fact that a time point or interval can be expressed by an explicit time indicator, or by an occurrence.

The same is true for spatial expressions:

77 (CL) *The two boats ran into each other where the river flows into the sea.*

78 (IC) *The two boats ran into each other near the delta of the river.*

and the manner in which an occurrence unfolds:

79 (CL) *He draws as his instructor told him.*

80 (IC) *He draws beautifully.*

The idea that certain types of concepts can have different expressions is presented in (Thompson and Longacre, 1985). The authors identify clauses that can be substituted by a single word. In particular, adverbs can replace clauses expressing time, space or manner:

> ... What we are claiming is that the *semantic relationship* between the adverbial clause and the main clause is exactly the same as that between the adverbial word and the main clause. That is, either a (non-anaphoric) word or an entire clause can express the time, locative, and manner relationships. (178)

Thompson and Longacre perform an analysis across several languages (Isthmus Zapotec from Mexico, Barai from Papua New Guinea, Hausa, Mandarin, etc) to show that this is a pervasive phenomenon in language, not limited to the particulars of English.

3.4.3.3 Deletion

Sometimes the verb is used only to link two of its arguments. If a link can be inferred without the verb, then the verb can be deleted, like in the example:

81 (IC) *That virus causes the flu.*↔(NP) *flu virus (*Cause*)*

cause is a specific predicate that imposes strong constraints on the type of arguments and the link between them. There are, however, other verbs, like *be*, which are very general, and do not bring much semantic information to the expression. They can be deleted as well, without loss of information:

82 (IC) *The dog is brown.* ↔(NP) *brown dog (*PROPERTY*)*

83 (IC) *The printer is in the lab.* ↔(NP) *lab printer (*LOCATION*)*

The constraints on the arguments of *be* and *have* are looser than the ones imposed by semantically stronger predicates like *cause*. As seen in the examples for the verb *be* presented above, different relations can describe the interaction between arguments linked by the same predicate.

Levi (1978) proposes a list of nine *recoverable deletable predicates* (RDPs), five of which are verbs – *cause, have, make, use, be*[5]. By deleting one of these verb predicates, the expression becomes a noun phrase with similar meaning, for example:

84 (IC) *The man has a beard.* ↔(NP) *bearded man*

Levi has designed a set of transformations that systematically transform an expression into a semantically equivalent complex nominal, by deleting one of the RDPs. The transformations proposed are reversible but not easily automatable.

Peterson (1985), in an analysis of causative verbs and their paraphrases, claims that the verb *cause* used in a sentence has only the role of connecting the *Agent* with the embedded clause. It would be misleading to say that the *Agent* in the sentence is in fact the *Agent* of the action expressed by *cause*, like in the example:

85 *John caused the floor to be black. (*as a paraphrase of: *John painted the floor black.)*

The explicit use of the verb *cause* expresses the *Agent's* (John) relation to *the floor being black*. In some cases, such predicates can be omitted, like in the following example:

86 (IC) *virus causes flu* ↔(NP) *flu virus*

Deletion is similar to *metonymy*. In both cases a part of the occurrence does not find a counterpart in a surface expression. The difference between the two phenomena comes

[5] The other ones are *in, for, from, about.*

from the fact that *deletion* is an "internal" phenomenon – there is a combination of participants and attributes of the occurrence that can express the whole occurrence without the need for an overt expression of the concept core – while *metonymy* is "external" – it is the interaction with another concept that allows the part to conjure up the whole in the mind of the listener.

3.4.3.4 Word Formation

When an occurrence surfaces in a syntactic form, it can appear as a clause or just as a word. We have seen that metonymy covers such a case, when part of an occurrence will stand for the whole. The part may be the *Object* or the *Agent*, and it is usually expressed by a noun. Another mechanism to allow for the expression of an occurrence through a word is to allow the main verb of the action to surface in some form. It may surface as the head noun through nominalization:

87 *The investigator's* **report** *was very brief.*

or as a modifier through adjectivalization:

88 *The* **sunken** *ship was finally discovered.*

Word formation processes compact a whole expression in only one word, allowing a concept to take the form of various syntactic structures.

In the process of word formation, concepts expressed using a structure having a word in a certain open class as a head can also be expressed through a word in a different part of speech. The transformations from one class to another can be done through affixation (prefixation, suffixation), conversion, compounds or other modes. Quirk et al. (1985) present a comprehensive overview of methods of word formation. In the sections that follow we will give some example for each of the open classes – noun, verb, adjective and adverb.

The phenomena of word formation reveal to us another level in language – the morphological level. We assign semantic relations to pairs of concepts behind structures in texts, but how about relations between pairs of concepts behind one word? A *teacher* is the Agent of a teaching occurrence, a *hammer* is an Instrument in a hammering occurrence, etc. We have briefly explored this aspect (Nastase and Szpakowicz, 2003a). We leave further exploration of the morphological level to future work.

3.4.3.4.1 Nominal Expressions In relations that involve an occurrence and one of its participants, the occurrence is expressed by the main verb. The verb must be there, in some form. Nominalization is one of the phenomena that can account for the presence of a verb in a noun phrase.

For example:

89 (IC) *The students protested against tuition fee increase.*

can be almost equivalently expressed:

90 (NP) *student protest against tuition fee increase*

The main verb is expressed using a nominal form, and the whole clause is changed into a noun phrase.

We could expand on example (67) discussed above, repeated here as (91):

91 (IC) *The student was anxious because of the exam.*

This can be changed into the noun phrase *exam anxiety* by nominalizing the *Effect* proposition, which is expressed through a state-type occurrence. The nouns *exam* and *anxiety* are connected by causality, the information about who is in a state of anxiety has been omitted.

We include the genitive case in this discussion of nominal expressions. Nouns inflected for the genitive case can be paraphrases of, or be paraphrased as, clauses displaying different types of semantic relations (Quirk et al., 1985):

- Possession

  92 *Mrs. Johnson's passport ↔Mrs. Johnson has a passport.*

- Agent

  93 *the boy's application ↔The boy applied for ...*

- Object

  94 *the boy's release ↔... released the boy.*

- Measure

  95 *ten days' absence ↔The absence lasted ten days/ [She] was absent for ten days.*

etc.

Similarity between a clause and its nominalized version is largely recognized in linguistics and NLP, as it will become apparent from the following literature review.

Quirk et al. (1985) state that the structures of a clause and a noun-phrase can be mapped onto each other. They also distinguish several degrees of nominalization – ranging from gerunds to verbal nouns to deverbal nouns – all of which may retain the original verb's syntactic structure, as in the examples:

96 (IC) *The reviewers criticized his play in a hostile manner.*

97 (NP) *the reviewers' hostile criticizing of his play*

98 (NP) *the reviewers' hostile criticism of his play*

Macleod et al. (1998) are concerned with the correspondence between the arguments of a nominal form of a verb and its arguments in the verb form. Their concern is purely grammatical, with no interest in the semantic relations between the words, and they only consider the arguments that a verb subcategorizes for. They are concerned with transformations that relate the subject, object and indirect object to the arguments of the nominal form of the verb, and their relative positions in the new NP structure. They produce a nominalization lexicon that will contain the nominalization, the original verb, the verb's argument structure and the NP's argument structure.

In a continuation of this work, Meyers et al. (1998) describe the use of the nominalization lexicon in information extraction. An interesting part of this work is not just the use of patterns stored in the lexicon, but also semantic generalization. The example described is the following: reformulation according to the nominalization patterns of the sentence *IBM appointed Alice Smith.* Besides generating paraphrases, the program (PET) also generalizes *IBM* to *company* and *Alice* to *person*, generating the following nominalization pattern:

np(C-company)'s appointment of np(C-person)

(IBM's appointment of Alice Smith).

Optional temporal arguments are also considered, but since they do not interfere with the nominalization pattern, the authors say, they are not described in the paper.

In (Hull and Gomez, 1996), the focus is on differentiating between verbal and non-verbal senses of a nominalization (e.g. promotion, decoration, etc.), disambiguating the verbal sense of the nominalization (e.g. promotion of Peter, promotion of liberalism, etc.), and then assigning proper thematic roles to the modifiers of the nominal, according to the underlying verb. For each sense of a verb that appears in *WordNet*, there is a structure that identifies what (ontological) classes its arguments belong to. There are specific arguments that verbs require. Based on this fact, and on additional information that resolves the mapping between the nominalization's modifiers and verb arguments, one can decide whether the current sense is verbal. The same structure, together with the mapping between modifiers and arguments, is used for nominalization, to fill its thematic roles. *WordNet* is the ontology used to classify a verb's arguments, and also to determine in a first step whether the noun is a potential nominalization by checking if any of its senses are hyponyms of actions or events.

Rappaport-Hovav and Levin (1992) look at a group of nominals ending in *-er* (writer, bake, teacher), often called *agentive* nominals, and at the claim that they inherit the argument structure of the verb they are derived from. The authors look at different aspects of *-er nominals*, and conclude that the characteristic of nominals of inheriting the argument structure of the verb divides them into two classes – agentive and instrumental.

3.4.3.4.2 Adjectival Expressions In the case of verb nominalization, the transformation can affect the arguments of the verb as well, like in the example:

99 (IC) *The parents refused ...*

which can be equivalently expressed through a noun phrase:

100 (NP) *parental refusal*

In this case the *Agent* of the occurrence has been adjectivalized, while the head verb has been nominalized.

Adjectivalization can also occur in conjunction with deletion:

101 (IC) *The nation has a large debt.* ↔(NP) *national debt*

The verb can be used in its past participle form as a modifier (deverbal adjective):

102 (NP) *vanished treasure*

There are then two different phenomena of interest: the adjectivalization of the noun, and the adjectivalization of the verb.

The participle form of a verb can be used as an adjective. There are adjectives that are not derived from verbs, although they have the same suffix as participles (for example *talented, diseased, unexpected*). We are not concerned with them here.

According to Quirk et al. (1985), participial adjectives with the suffix *-ed* that have a corresponding verb usually have a passive meaning, except in the case when the original verb is intransitive. The attributive form corresponds then to a passive paraphrase:

103 *lost property* ↔*property that has been lost*

If the participle corresponds to an intransitive verb (or the intransitive use of a verb), there is no passive interpretation, but there is a corresponding clausal paraphrase:

104 *the escaped prisoner* ↔*the prisoner that escaped*

105 *the grown boy* ↔*the boy who has grown*

In an analysis of non-predicating adjectives, Levi (1978), claims that denominal adjectives inherit the thematic roles of the original noun. Raskin and Nirenburg (1995) also support this view:

agentive: *presidential refusal, editorial comment*

objective: *constitutional amendment, oceanic study*

locative: *marginal note, marine life*

dative[genitive]/possessive: *feminine intuition, occupational hazard*

instrumental: *manual labour, solar generator*

They consider that this property is not relevant in describing adjectives. We consider it interesting though, since it allows us to show instances of the same relation on different syntactic levels.

3.4.3.4.3 Verb Expressions Verbs can be formed from nouns and adjectives.

Denominal verbs have an implicit case. One of the entities in this case relation is the noun at the origin of the denominal verb, and the other is an unspecified action.

Examples:

hammer – implicit case Instrument

106 (IC) *We hammered the nail into the board.*

It is implied that a *hammer* was used in the most common manner to hit something. Then *hammer* is actually the *Instrument* in some unspecified action.

tape – implicit case OBJECT

107 (IC) *We have taped the whole meeting.*

This means we have used some *tape* in the most common way it is used when it is involved in an event that is to record onto it. So the original noun *tape* would be the *Object* of *record* which is implied.

The *most common way in which something is used* can be obtained through statistical analysis of a large corpus, or from dictionary definitions. Word sense disambiguation is also necessary to distinguish the correct sense of *hammer* and *tape* and the verbs with which they are associated in the corpus.

According to Kelly (1998), there are two types of denominal verbs.

- **rule-derived** (RD). The sense of the verb can be inferred from the semantic category of the original noun. The relation between the verb and the noun is captured in word definitions. Example:

  108 *We hammered the two boards together.*

  to hammer

  - *LDOCE* (1978): (sense 1) to hit something with a hammer in order to force it into a particular position or shape.
  - *Collins* (1986): (sense 12) to strike or beat (a nail, wood, etc.) with or as if with a hammer.
  - *WordNet 1.6*: (sense 1) beat with or as if with a hammer.

- **idiosyncratically derived** (ID). The sense depends on idiosyncratic aspects of the original noun. Example:

  109 *He pigged out at the buffet.*

  We understand that *to pig out* means *to eat in a certain manner, like pigs do.* The relation between the verb and the noun is not explicit in the verb definition:

- *LDOCE* (1978): (sense 1) to eat a lot of food.
- *Collins* (1986): (sense 1) to gorge oneself.
- *WordNet 1.6*: (sense 2) eat greedily.

For our purpose of detecting semantic relations, the second type of denominal verbs are a more concise expression of a longer paraphrase. The paraphrase is replaced by a word because of some analogy between a characteristic associated with an entity expressed by a noun, and the ideas conveyed by the paraphrase. Without the appropriate knowledge resource, such denominal verbs are hard to analyze.

The rule-derived denominal verbs however, are intrinsically connected to the original noun. This is interesting for our work. It means that the verb expresses an occurrence in which the entity denoted by the noun plays some role, which we want to identify.

Verbs can also be derived from adjectives:

110 *They cleaned the room.*

3.4.3.4.4 Adverbial Expressions Adverbs can be formed from adjectives by suffixation:

slow ↔*slowly*

Adverbs expressing direction can be formed by adding to the noun that shows the direction the suffix -*ward*:

toward home ↔*homeward*

toward north ↔*northward*

Temporal expressions can also be compacted and expressed through an adverb:

every week ↔*weekly*

every day ↔*daily*

3.4.3.4.5 Pronouns Although they are closed class words, pronouns can affect the structure of the sentence. More specifically, possessive pronouns can replace a structure centered on a verb expressing possession, like *have* or *belong to*, as in the example:

111 *She has a book.* ↔*her book*

112 *The book belongs to her.* ↔*her book*

The relation between *book* and *she* is Possession in all these cases. The emphasis changes between the expressions – *book* is emphasized in the second expression as opposed to the first one, when *she* is the one in focus.

3.5 Using the Phenomena Identified

The phenomena that account for different surface forms of concepts serve first of all as evidence for the validity of the unified view of semantic relations across syntactic levels.

By systematically relating all possible surface forms of the same concept, we explain why a concept can have expressions that pertain to different syntactic levels. If an occurrence can be expressed through a clause, as well as a noun-phrase, a relation that links two occurrences can surface at the clause, intra-clause, or even noun-phrase level. Because of the fact that concepts can surface through various forms, and semantic relations link concepts, we conclude that semantic relations are in fact the same across syntactic levels, and they surface on the level on which the concepts they connect surface.

This is explored and exemplified in Chapter 4, where we take semantic relations found in the literature, we explore the types of concepts they connect, we justify the surface forms these concepts may take using the phenomena discussed in Section 3.4.3, and we show the various syntactic levels on which these relations can be found.

We also use these phenomena to unify the three separate lists of relations presented in Section 1.4, and as indicators in a semi-automatic text analysis and knowledge acquisition system.

As it was briefly presented earlier, the system that is the goal of this research project uses various syntactic and semantic information to help in the task of semi-automatically assigning semantic relations to pairs of entities, occurrences, attributes or parts of occurrences. The linguistic phenomena presented in the preceding sections give the system some of the desired clues. Recognizing that an expression was the result of a certain process of word formation should provide good indication of the underlying semantic relation.

The clues given by nominalizations, adjectivalizations and so on, are very important. They are indicators of a morpho-syntactic nature, which in some cases are the attributes that give the best rule to characterize a relation. This is the case for one of the relations we work with, Object-Property. At the noun-phrase level it is unambiguously characterized by the following rule:

If the modifier is a past participle, then the semantic relation between the two entities is Object-Property.

113 (IC) *The ship sank* ⟵(NP) *sunken ship*

114 (IC) *They repaired the engine* ⟵(NP) *repaired engine*

115 (IC) *The treasure vanished* ⟵(NP) *vanished treasure*

According to the type of their best indicators, we consider that the semantic relations vary along a syntactic-semantic axis. The relation Object-Property is located toward the syntactic end of the spectrum. The relations are spread along this axis, and the less informative the syntactic indicators for a certain relation, the harder it is to find clear cut rules to define it, based solely on morpho-syntactic indicators. But morpho-syntactic indicators definitely have an important role among the indicators we consider.

In order to establish a connection between indicators and semantic relations, we will use this information in machine learning experiments designed to find rules for semantic relation assignment based on syntactic and semantic indicators. These experiments are described in Section 6.3.

3.6 Across Languages

We consider the question whether concepts can surface in different syntactic forms in languages other than English, and if the phenomena we have identified in the previous sections account for such forms across languages. We choose a few languages from the same Indo-European family, because of Whorf's principle of language relativity (Whorf, 1956). His examples and discussion of American-Indian languages have shown that there are languages in which concepts are different than the ones people within the western culture are used to deal with. We will therefore not stretch this discussion beyond the family in which the English language is included. *Equivalence* will be an exception, since there are studies to support this phenomenon in various exotic languages.

Among the phenomena proposed, metonymy, equivalence and deletion seem to be rather language independent, and more conceptual in nature. If that is the case, we should find proof of their existence in the form of examples similar to those in English that gave us clues about their existence in the first place.

Word formation processes are more language-dependent, but they appear in all languages. The extent to which words from different parts of speech are related to one another through word formation phenomena is expected to vary.

We choose to present below examples for each of the phenomena we have identified in a few languages from the Indo-European family: English (EN), Russian (RU), Italian (IT) and Romanian (RO). These examples were chosen so that the expressions are equivalent in the languages presented.

Metonymy

EN 116 *Sweets before dinner spoil your appetite.*

117 *Eating sweets before eating dinner spoil your appetite.*

RU 118 Сладости перед обедом портят аппетит.

sweets before dinner spoil appetite

119 Если вы едите сладости перед обедом, вы портите свой аппетит.

if you eat sweets before dinner, you spoil of-self$_{2sg}$[6] appetite

IT 120 *Dolci prima della cena rovinano l'appetito .*

sweets before of-the dinner spoil$_{3pl}$ the appetite

121 *Se mangi dolci prima di cennare ti rovini l'appetito.*

if eat$_{2sg}$ sweets before of having-dinner to-you spoil$_{2sg}$ the appetite

RO 122 *Dulciuri înainte de cină strică apetitul.*

sweets before of dinner spoil appetite$_{def.art.}$[7]

123 *Dacă manânci dulciuri înainte de a cina îţi strici apetitul.*

if eat$_{2sg}$ sweets before of to dine to-you spoil$_{2sg}$ appetite$_{def.art}$

Examples (116), (118), (120) and (122) are equivalent. Examples (119), (121) and (123) are slightly different than their English counterpart, and they can be literally translated as:

124 *If you eat sweets before having dinner you spoil your appetite.*

The change is due to the fact that a literal translation of example (117) did not sound natural in Russian, Italian and Romanian. The phenomenon of metonymy, however, is clearly captured.

Equivalence

[6] 2sg = second person singular.

[7] def.art. = definite article

EN 125 *They practice while students have lunch.*

126 *They practice during lunch hour.*

RU 127 Они репетируют когда студенты едят обед.

they practice when students eat lunch

128 Они репетируют во время обеденного перерыва.

they practice in time lunch$_{adj.}$ break$_{GEN}$[8]

IT 129 *Loro studiano mentre gli studenti mangiano il pranzo.*

they practice$_{3pl}$ while the students eat the lunch

130 *Loro studiano durante l'ora del pranzo.*

they practice$_{3pl}$ during the hour of-the lunch

RO 131 *Ei exersează în timp ce studenţii iau prânzul.*

they practice$_{3pl}$ in time of students$_{def.art.}$ take lunch$_{def.art.}$

132 *Ei exersează în timpul orei de prânz.*

they practice in time$_{def.art.}$ hour$_{GEN}$ of lunch

Examples (125), (127), (129) and (131) are equivalent, and so are examples (126), (128), (129) and (132).

Thompson and Longacre (1985) present examples in more exotic languages (Hausa, Mandarin, etc.) that allow for temporal, locative and manner clauses to be substituted by single words, while the semantic relations remain the same. We reproduce a few examples from (Thompson and Longacre, 1985), in Isthmus Zapotec, an Otomanguean language of Mexico (the italics emphasize the clauses and words that express temporal, spatial and manner relations; the emphasis is the authors'):

Time :

a.	Kundubi	bi	*yánaji*	
	is blowing	wind	today	
	'It's windy today'			
b.	*Ora*	*geeda-be*	zune	ni.
	when	(POT)come-he	(FUT)do I	it
	'When he comes I'll do it'			

[8] GEN = genitive case

Locative :

a.	Nabeza	Juan	*rarí*		
	dwells	John	here		
	‘John lives here’				

b.	*Ra*	*zeeda-be-ke*	nuu	ti	dani
	where	is coming-he-that	is	a	hill
	‘Where he was coming along, there was a hill’				

Manner :

a.	*Nageenda*	biluže-be
	quickly	finished-he
	‘He finished quickly’	

b.	Gu’nu	*sika ma*	*guti-lu*
	(POT)do you	like already	(COMPL)die-you
	‘Act as if you’re dead’		

Deletion

English allows for easy forming of base noun phrases through the deletion of a predicate. Deletion is possible in other languages as well, but it does not produce as compact an expression as it does in English.

EN 133 *the machine that makes bread ↔bread-maker*

RU 134 машина которая выпекает хлеб ↔автомат для выпечки хлеба

machine that bakes bread ↔automaton for baking$_{dev.n.}$[9] bread$_{GEN}$

IT 135 *la machina che fa il pane ↔machina da pane*

the machine that makes the bread ↔machine for bread

RO 136 *maşina care face pâine ↔maşina pentru pâine*

the machine that makes bread ↔machine for bread

Word Formation

Other languages form more easily nouns to name entities according to their properties, which English does not allow:

[9] dev.n. = deverbal noun

EN 137 *the hanged man* ↔*the hanged man*

RU 138 повешенный человек ↔повешенный

hanged human ↔hanged$_{masc.sg}$[10]

IT 139 *l'uomo impiccato* ↔*l'impiccato*

the man hanged ↔the hanged$_{masc.sg.}$

RO 140 *omul spânzurat* ↔*spânzuratul*

man$_{def.art.}$ hanged ↔hanged$_{def.art.masc.sg}$

Whereas in English we may say *the hanged*, this noun will denote a class of people as opposed to a specific individual.

The examples illustrate nominalization in which a noun-adjective compound is nominalized, but the core of the nominalization is the adjective not the head noun.

Here is an instance of verb nominalization, across the languages we look at:

EN 141 *The students protested against tuition fee increase.*

142 ↔*student protest against tuition fee increase*

RU 143 Студенты протестовали против повышения оплаты за обучение.

students protested against increase$_{GEN}$ pay$_{GEN}$ for tuition

144 студенческий протест против повышения оплаты за обучение

student$_{adj.}$ protest against increase$_{GEN}$ pay$_{GEN}$ for tuition

IT 145 *Li studenti hanno protestato contro il aumento delle tasse scolastiche.*

the students have protested against the increase of-the fees tuition$_{adj.}$

146 *protesto degli studenti contro l'aumento delle tasse scolastiche*

protest of-the students against the increase of-the fees tuition$_{adj.}$

[10] masc.sg. = masculin singular

RO 147 *Studenţii au protestat împotriva creşterii taxelor de şcolarizare.*

students$_{def.art.}$ have protested against increasing fees$_{GEN}$ of tuition

148 *protestul studenţilor împotriva creşterii taxelor de şcolarizare*

protest students$_{GEN}$ against increasing fees$_{GEN}$ of tuition

Nominalization is not always as successful in compressing expressions in languages other than English:

EN 149 *The student was anxious because of the exam.* ↔*exam anxiety*

IT 150 *Il studente era ansioso a causa del esame.* ↔*ansieta a causa del esame*

the student was anxious of cause the exam ↔anxiety of cause of-the exam

RO 151 *Studentul era agitat din cauza examenului.* ↔*agitaţie din cauza examenului*

student$_{def.art.}$ was anxious of cause exam$_{GEN}$ ↔anxiety of cause exam$_{GEN}$

They are all noun phrases, but not base NPs like in English. We observe in the examples (149) to (151) that whereas nominalization of the state of *anxiety* is possible in all languages, the predicate indicating causality could not be deleted from the phrases in Italian or Romanian as easily as it was deleted from the English version. This observation suggests an interesting experiment, should multilingual corpora be available. In a parallel multilingual corpora, we could find parallel expressions of the same concepts in different languages. Some languages need to keep some indicators of the relation between the concepts presented in the utterance (as is the case for *causa* (IT), and *cauza* (RO)). The parallel expressions that are richer in such information could then be used to assign semantic relations to the other expressions.

3.7 Conclusions

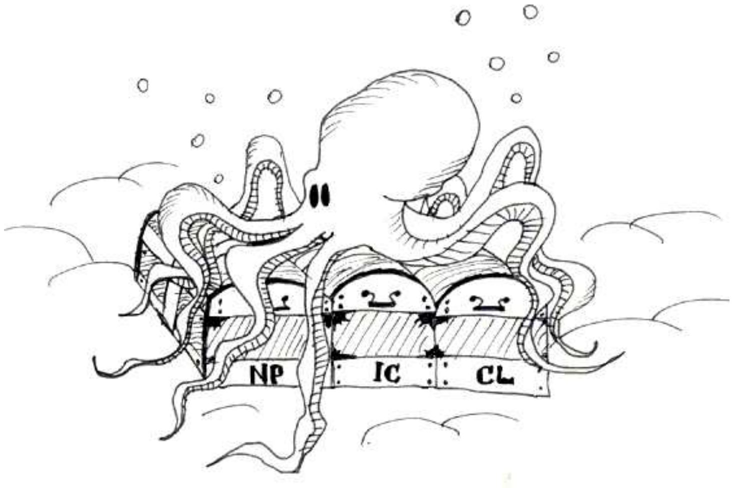
We have explored in this chapter the idea of expressing the same concept through different syntactic forms.

We have looked at various syntactic structures that convey very similar meaning (they express the same concepts that interact in the same way). We have identified linguistic phenomena that justify the existence of various forms that can convey very similar meaning. Some of the phenomena identified are conceptual, and depend less on the particular language under analysis, as the very brief survey across languages has shown in Section 3.6. Although the process of word formation is language-dependent, it does have counterparts in the various languages we have looked at.

The phenomena we have identified and analyzed in this chapter constitute the proof that semantic relations are the same across syntactic levels. We will encounter the semantic relations on any level on which the concepts they connect can surface.

Chapter 4

Semantic Relations Across Syntactic Levels



4.1 Introduction

In the previous chapters we have seen that the same idea can surface in different syntactic forms. We are therefore led to reconsider the semantic analysis of a text from the point of view of relations that describe interaction between pairs of concepts, rather than syntactic units. If the same pair of concepts underlies pairs of different syntactic units, the links between the concepts' different surface expressions should be the same.

From this perspective of connections between concepts that happen (because of certain factors, for example parsimony) to take one or another syntactic form, we have to analyze semantic relations in a different manner. We will start from the name of a semantic relation that was intuitively assigned to describe a certain phenomenon, and we will try to understand its meaning. This will lead us to understand what a relation implies in terms of describing the interaction/link between two concepts.

To take an example, let us consider a temporal relation, Time Through. This relation describes the relative position on the time axis of an occurrence and a time interval. The occurrence unfolds all through the specified time interval. But the time interval can be expressed in many ways. It can be a definite time interval (*three hours*), or it can be defined through another occurrence that satisfies the requirements of the relation (*the soprano sang an aria*) in this case an occurrence that is bound in time.

If the concepts whose interaction a semantic relation describes are the same, despite their different surface syntactic forms, the semantic relation is the same. Therefore we can find the same semantic relation on different syntactic levels.

Instead of grouping semantic relations by the syntactic level they appear on, we take semantic relations one by one, determine what type of concepts they link, and look for examples that display the various syntactic manifestations of such concepts. The semantic relations are then clustered by the type of interaction they describe.

We justify and present examples for a unified view of semantic relations, for relations in the following six classes: **causality**, **temporality**, **spatiality**, **conjunctive**, **participant**, **quality**. This grouping, as well as the relations analyzed, is not random. These classes arise from comprehensive research on the topic (Barker et al., 1997a). For each class of relations we will show how it appears in the literature, and how the relations that it includes are described.

Not all these relations have instances on all the syntactic levels we consider, and when this happens we will explain why. This fact has no negative effect on the unified view we propose. Our purpose is not to "force" semantic relations to have expressions at all syntactic levels that we consider. Instead, we look at what each semantic relation means and what type of concepts it links. Next, we study what surface forms these concepts can take, and we show on which syntactic levels each semantic relation can have instances.

Although we show particular semantic relations and the way they surface on different syntactic levels, the considerations for unification are general.

Each relation exemplified is accompanied by a definition and instances are shown for all the levels at which it can be found. To help the reader follow the syntactic phenomena responsible for the expression of similar concepts in different forms, the examples presented for each relation are kept as consistent as possible. By that we mean that the examples are chosen so that the same concepts underlie the expressions on different syntactic levels. In situations where this produced examples with questionable wording, or when we want to make a point about a certain type of surface expression, this constraint is disregarded.

4.2 Causality

4.2.1 General Considerations

Semantic relations in this class describe the causal interaction between two occurrences, one of which influences in some way (causes, opposes, enables, etc.) the other.

The idea that causality is a relation between two events was adopted from philosophical (Davidson, 1967) and psychological (Schank, 1973), (Miller and Johnson-Laird, 1976) research. In linguistics there are three views of this issue. Causality is regarded as a relation between an agent and an event (Jackendoff, 1990), (Rappaport and Levin, 1988), as a relation between two events (Dowty, 1979), or as a relation between two entities, an *Agent* and a *Patient* (Lakoff and Johnson, 1980). The first two views arise from the analysis of verb meanings.

For example, the sentence:

152 *He sweeps the floor clean.*

is differently analyzed under the two views. In the agent-event relation, the above sentence becomes (Rappaport and Levin, 1988):

x CAUSE [floor BECOME (AT) clean BY [x 'wipe' floor]]

or (Dowty, 1979)

[[He sweeps the floor] CAUSE [BECOME [*the floor is clean*]]]

The third view arises from an analysis of causation as a basic, but decomposable, concept of human activity (Lakoff and Johnson, 1980). Causality is decomposed in terms of separate actions that the *Agent* undertakes in order to achieve a change of state in the *Patient*.

Givon (1975) analyzes the causal construction in English on different syntactic levels. He starts from the premise that *cause* is a predicate that takes two sentential arguments, one to denote the *Cause* (P_c), and the other the *Effect* (P_e). Of these two arguments, according to the language expressions, the subject of P_c is the one usually considered to be the subject of the causative expression, and the subject of P_e is considered the object of the causation. In order to explain the different paraphrases of such a causal relation, Givon describes the *Raising/Foregrounding* process, which condenses the two sentential arguments P_c and P_e into a single proposition. The end result could be a sentence with only one lexical verb. The initiator and the undergoer will be the subject and object of the causative verb, respectively. Examples given in the article:

153 *George shot the gun at the elephant, and as a result, the elephant died.*
George shot the gun at the elephant, and thus caused the elephant to die.
George's shooting the gun at the elephant caused the elephant to die.
George caused the elephant to die by shooting the gun at him.
George caused the elephant's death by shooting the gun at him.
George killed the elephant by shooting the gun at him.
George killed the elephant with the gun.

Givon notes the following:

> By positing a "weak relatedness" between these examples, we do not wish to suggest that they are derivationally/transformationally relatable, nor that they share a considerable portion of their semantic structure. Paraphrases of this type are at best suggestive. [...] The examples are different from focus/topic point of view.

The transformation proposed for the example presented is the following:

P_c [cause] P_e
Nom_a [cause] P_e by P_c
Nom_a [cause-v_e] Nom_p by Nom_i

where Nom_a is the *Agent* of P_c, Nom_p is the *Patient*-subject of P_e, and v_e is the verb of P_e.

The view we propose is that causality relations hold between two occurrences. Nonetheless we agree with the idea that sometimes clause-level relations are better explained by considering certain constituents of the clauses other than the head verb. For some relations, the subject of P_c is not always the *Agent* of the occurrence that causes P_e, but it can also be the *Object*, as illustrated in the example:

154 *The file printed because the program issued a command.*

Because the sentence can be paraphrased into a sentence that has a subordinate clause:

155 *The file printed because of the command issued by the program.*

the semantic relation is not a clause-level relation between *printed* and *issued* but rather a relation between the event of *printing* and an unspecified event brought about by *command.* In this case, *command* is the *Agent* of some unspecified action that is the true cause of the file being printed.

We seek similarities between different syntactic realizations of Causality relations.

There is work that emphasizes the differences in such expressions. In (Peterson, 1985) there is a distinction between non-agentive and agentive causal expressions, like in the following examples:

Non-agentive

156 *The crack in the engine mount caused the engine to fall off.*

157 *The engine falling off caused the plane to crash.*

Agentive

158 *The pilot caused the plane to crash [by falling asleep].*

159 *The mechanic caused the engine to fall off.*

Both the *Cause* and *Effect* propositions in a causal relation are events, only not always fully specified. In the examples above, particularly in the agentive ones, we regard the *pilot* and the *mechanic*, respectively, as agents of some occurrence that is the actual *Cause.*

In example (158), it is the occurrence of the pilot falling asleep that causes the plane to crash, and in example (159) it is an unspecified action that the mechanic performed that caused the engine to fall off.

Peterson (1985) is concerned mostly with causative verbs, and alternative expressions of sentences that contain such verbs. The type of paraphrases presented (in terms of the verb *to cause*) are not similar to the phenomena that we are analyzing.

Barrière (2001) also analyzes causality and looks for patterns indicative of causal relations in texts. Some of these patterns (connectives – *because, cause, and, if-then*, etc. – and prepositions – *by, through*, etc.) are also indicators in the heuristics designed for the inter-clause and clause level. The best indicators for causal relations in (Barrière, ibid.) are verbs that express different types of causal relations (e.g. *increase, decrease, create, destroy, maintain*). In this research, lexical information about verbs that express causality is not explicitly used.

4.2.2 Causality Case Study

Relations grouped under **causality** are relations that describe how two occurrences influence each other. The concepts that these relations connect must be occurrences. These occurrences can be expressed by clauses, noun phrases or modifiers. From the phenomena that account for the different surface forms of occurrences, *metonymy, deletion* and *nominalizations* are the ones that manifest themselves in causal relations. Figure 4.1 represents, in an attribute-value format, the essence of the causal relations.

$$\left[\begin{array}{ll} \text{OCCURRENCE1} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{Occurrence1} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \text{occurrence part type} \\ \text{FILLER} & \text{occurrence part} \end{array}\right] \end{array}\right] \\ \text{OCCURRENCE2} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{Occurrence2} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \text{occurrence part type} \\ \text{FILLER} & \text{occurrence part} \end{array}\right] \end{array}\right] \\ \text{INDICATOR} & \text{indicator} \end{array}\right]$$

Figure 4.1: The essence of causal relations

Figure 4.2 shows an instantiation of the causal pattern for the EFFECT relation:

160 $\underline{\textit{The student was anxious}}_{\textit{effect}}$ *because* $\underline{\textit{he was writing an exam}}_{\textit{cause}}$.

We distinguish several causal relations. The differences among them come from the different aspects of the occurrences they connect. We will present each situation, and the syntactic levels on which they may be encountered.

4.2.3 Definitions

- CAUSE: 1 causes 2. 1 is sufficient to cause 2, and 1 is known to exist. As a consequence, 2 will exist.

$$\begin{bmatrix} \text{CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{write} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{Object} \\ \text{FILLER} & \text{exam} \end{bmatrix} \end{bmatrix} \\ \text{EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{be anxious} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{-} \\ \text{FILLER} & \text{-} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{because} \end{bmatrix}$$

Figure 4.2: Attribute-value representation for the EFFECT relation

161 (CL) *He went blind*$_2$ *because the snow was shining sharply*$_1$.

(IC) *He went blind*$_2$ *because of the snow*$_1$.

(NP) *snow*$_1$ *blindness*$_2$

- EFFECT: 2 is the result of 1. (1 causes 2). 2 is the focus

162 (CL) *The program issued a command*$_1$ *so the file printed*$_2$.

(IC) *The command*$_1$ *caused the printing of the file*$_2$.

(NP) *print*$_2$ *command*$_1$

- PURPOSE: 1 is for 2, but 2 does not necessarily come into being.

163 (CL) *She took the medicine*$_1$ *so the pain should be relieved*$_2$.

(IC) *She took the medicine*$_1$ *for pain relief*$_2$.

(NP) *pain-relief*$_2$ *medicine*$_1$

- ENTAILMENT: 1 entails 2. 1 is not known to exist or not, but if it does then necessarily 2 also exists.

164 (CL) *If students work hard*$_1$, *they pass their exams*$_2$.

(IC) *Hard-working students*$_1$ *pass their exams*$_2$.

(NP) *a pass*$_2$ *due to hard work*$_1$

- ENABLEMENT: 1 enables 2. 1 is necessary but not sufficient to make 2 exist.

165 (CL) *The printer can print*$_2$ *if the paper tray is full*$_1$.

(IC) *The printer can print*$_2$ *from a full paper tray*$_1$.

(NP) no instance found

- DETRACTION: 1 detracts/opposes 2, but the existence of 1 may not be sufficient to prevent 2 from existing.

166 (CL) *They persisted*$_2$ *although I warned them*$_1$.
(IC) *They persisted*$_2$ *despite my warning*$_1$.
(NP) *persistence*$_2$ *despite warnings*$_1$

- PREVENTION: 1 prevents 2. If 1 is known to exist, then 2 necessarily does not exist.

167 (CL) *The service did not work* $_2$ *since the hard-disk crashed*$_1$.
(IC) *The service did not work* $_2$ *because of a hard-disk crash*$_1$.
(NP) *service breakdown*$_2$ *on account of a crash*$_1$

4.3 Temporality

4.3.1 General Considerations

This class contains relations that describe the position of an occurrence on the time axis relative to a time expression. A point or interval of time can be expressed by an explicit time, or through an occurrence that can express a point in time (*the clock struck midnight*) or a time interval (*he filled the bottle with tap water*) introduced by a marker that has a temporal reading in that particular context (*when, since, etc.*).

In the literature on temporal relations, the focus is on the representation of the flow of events in order to facilitate certain tasks, and not on the analysis of linguistic phenomena for the discovery of temporal relations among concepts.

The aspects of the relations that are considered interesting are deeper than the ones we are looking for. The general focus is on representing and arranging in a temporal order occurrences detected in texts. We focus on smaller pieces of text, and the relations we assign are located inside a sentence. We will not go into the deep meaning of words to decide whether a certain occurrence still holds while another may hold or may have finished already. We try to use surface indicators and available information from lexical resources to distinguish between possible temporal relations. These can be further analyzed, according to the requirements of the task at hand.

Allen (1984) introduces a theory of temporal relations and representation of events, to address issues concerning different types of occurrences: actions that involve non-activity, actions that cannot be defined by decomposition into sub-actions and actions that occur simultaneously and may interact. This analysis of temporality revolves around the analysis of events and their relative position on the time axis, using the temporal aspects that the main verb of the event provides. We focus on positioning occurrences on the time axis relative to time indicators in the sentence, rather than analyzing the verb for particular temporal attributes.

Larson (1998) looks at modifiers that fill a temporal role, and underlines differences in interpretation. Prenominal modifiers are in the domain of a generic quantifier, postnominal modifiers are not. For example:

the lecture Thursday – *Thursday* denotes a specific Thursday,
the Thursday lecture – *Thursday* usually denotes a generic Thursday

In our **Temporal** relations, the first example would be assigned a TimeAt relation, indicating a specific time when an occurrence takes place, while the second one would be Frequency, indicating a repeating occurrence (in this case a lecture that takes place every Thursday).

The difference between the interpretation of pre- and post-nominal modifiers is more general than the case when the modifier expresses time. Bolinger (1967) arrived at the following conclusions:

> The post-nominal adjective attributes a temporary property.
>
> A prenominal adjective can attribute a temporary property, but also a characteristic or enduring property.

In cases other than the temporal example presented above, our list of relations does not distinguish between temporary or enduring aspects of the modifiers.

Miller and Johnson-Laird (1976) observe that all major classes of words can be used to express temporality (verbs: *end, precede*, nouns: *day, month, tomorrow*, adjectives: *former, present, successive*, adverbs: *eventually, often, soon*, prepositions: *at, during*, conjunctions: *as soon as, before, until*), besides the tense, which is explicitly represented in the form of the verb. Some temporal marker may have different functions, and introduce different types of temporal expressions, as shown in the following sentences:

168 *It happened <u>before noon</u>.*

169 *It happened <u>before he left</u>.*

In the first sentence, *before* is a prepositon that introduces a prepositional phrase which has an adverbial function. In the second example, *before* is a conjunction that connects two clauses. The semantic relation is the same.

4.3.2 Temporality Case Study

All relations grouped in the **Temporality** class share the following property: one of the entities they connect is an occurrence, the other serves as a reference on the time axis. The reference may be direct, through an explicit time interval, or indirect, through another occurrence. There may be also the case, for the TIMEAT relation, that both entities linked are time references. *Equivalence* and *word formation* are the phenomena that allow for a temporal concept to surface in different syntactic forms.

Figure 4.3 shows a representation in the attribute-value format that captures the essence of temporal relations.

$$\begin{bmatrix} \text{TIME1} & \begin{bmatrix} \text{TYPE} & \text{type of TIME1 indicator} \\ \text{FILLER} & \text{TIME1 indicator} \end{bmatrix} \\ \text{TIME2} & \begin{bmatrix} \text{TYPE} & \text{type of TIME2 indicator} \\ \text{FILLER} & \text{TIME2 indicator} \end{bmatrix} \\ \text{INDICATOR} & \text{indicator} \end{bmatrix}$$

Figure 4.3: The essence of temporal relations

The time indicator can be an occurrence or a definite time expression. Figure 4.4 shows an instantiation of the temporal representation for the TIMEAT relation for the noun phrase:

170 $\underline{\text{summer}}_{TimeAt}$ $\underline{\text{semester}}_{entity}$

4.3.3 Definitions

- CO-OCCURRENCE: 1 and 2 occur or exist at the same time. 1 and 2 express unbounded time intervals, they both represent occurrences.

  171 (CL) $\underline{\textit{He writes novels}}_1$ *while* $\underline{\textit{he listens to music}}_2$.
  (IC) $\underline{\textit{He writes novels}}_1$ *while* $\underline{\textit{listening to music}}_2$.
  (NP) $\underline{\textit{writing novels}}_1$ *while* $\underline{\textit{listening to music}}_2$

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{semester} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME AT} & \begin{bmatrix} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{summer} \end{bmatrix} \\ \text{INDICATOR} & - \end{bmatrix}$$

Figure 4.4: Attribute-value representation for the TimeAt relation

- Frequency: 1 occurs every time 2 occurs. 1 is an occurrence, 2 can be an occurrence or a point or time interval that appears several times on the time axis.

172 (CL) *We play volleyball*$_1$ *every time he visits*$_2$.
(IC) *We play volleyball*$_1$ *every week*$_2$.
(NP) *weekly*$_2$ *game*$_1$

- Precedence: 1 occurs or exists (or begins to occur or exist) before 2. Either 1 or 2 is an occurrence, the other can be an occurrence or an explicit temporal expression.

173 (CL) *I watered the flowers*$_1$ *before I left for the holidays*$_2$.
(IC) *I watered the flowers*$_1$ *before leaving for the holidays*$_2$.
(IC) *watered the flowers*$_1$ *before 8 o'clock in the morning*$_2$.
(NP) *watering the flowers*$_1$ *before leaving for the holidays*$_2$

- TimeAt: 1 occurs when 2 occurs. Both 1 and 2 can be occurrences or explicit temporal expressions.

174 (CL) *He traveled there*$_1$ *when they called him*$_2$.
(IC) *He traveled there*$_1$ *last year*$_2$.
(NP) *winter*$_2$ *travel*$_1$

- TimeFrom: 1 began to occur when 2 occurred. 1 is an occurrence, 2 can be an occurrence or point/interval in time, but is considered punctual.

175 (CL) *He has been playing well*$_1$ *since we coached him*$_2$.
(IC) *He has been playing well*$_1$ *since January*$_2$.
(NP) *playing well*$_1$ *since January*$_2$

- TimeThrough: 1 existed while 2 existed. 1 is an occurrence, 2 can be either an occurrence that delimits an interval of time, or an explicit time interval.

176 (CL) *The band practices*$_1$ *while other students have lunch*$_2$.
(IC) *The band practices*$_1$ *during lunch hour*$_2$.
(NP) *lunch-hour*$_2$ *practice*$_1$

- TimeTo: 1 existed until 2 started to exist or occur. 1 is an occurrence, 2 can be an occurrence that is considered punctual, or a point in time.

177 (CL) *They partied*$_1$ *until their mother sent them to bed*$_2$.
(IC) *They partied*$_1$ *until 9 o'clock*$_2$.
(NP) *party*$_1$ *until dawn*$_2$

4.4 Spatiality

4.4.1 General Considerations

The relations describe the relative position of an occurrence or an entity with respect to a point, area or volume in the three-dimensional space. The location in space can be expressed by an explicit spatial location (*near the river*) or by an occurrence introduced by a spatial preposition (*where the two rivers meet*).

Spatial relations pose an interesting problem for list unification. One can regard them as not fit for unification because at the clause level they are roles, but at the noun phrase level this is not necessarily true, as shown in the following examples:

178 (IC) *He paints in the garden.*

179 (NP) *painting in the garden [is his hobby]*

180 (NP) *the painting in the garden [was ruined by the rain]* (the sense of *painting* here is *picture*)

In examples (178) and (179) the message conveyed is about an occurrence of painting that takes place in the garden. In both these examples *paint/painting* has a verbal reading, indicating an action. In example (179) the action is expressed through a gerund, but the semantic relation between the occurrence and the location is the same – LocationAt. It can be argued that example (180) is different, since *painting* is an entity, and the *garden* is now the location of an entity and not an occurrence, and therefore it is not the same type of relation.

We extend the notion of locality to encompass both the location of an occurrence and that of an entity. We see no reason why this should not be allowed – the expression of a **Spatial** relation inside a noun phrase whose head is a true noun (in the sense that it does not have a verbal reading) can be regarded as an expression from which the verb was omitted. For LOCATIONAT the verb may be the existential verb *to be*, as exemplified by the following phrase and its corresponding paraphrase:

181 (NP) *the painting in the garden*

182 ↔(CL) *The painting is in the garden.*

One could also regard the problem of having an entity-location relation as a specialized situation of the more general occurrence-location relation. In the occurrence-location relation, all the constituents of the occurrence will be involved in the same relation. For example:

183 (IC) *I ran through the valley.*

The *run* occurrence takes place through the location – *valley* – which means that the *Agent* of the occurrence will be located at different points in the valley at different points in time.

Miller and Johnson-Laird (1976) make a similar point in their analysis of spatial relations. In the sentence:

184 *The plumber worked in the kitchen.*

the adverbial phrase *in the kitchen* shows where the *work* occurrence takes place, but since the *Agent* must be located at the same place, one can say that locating the occurrence is equivalent to locating the *worker*, which in this case is the *Agent*. Then the following expressions: *The plumber worked in the kitchen*, *The plumber was in the kitchen*, and *the plumber in the kitchen* are all instances of the same spatial relation.

In (EAGLES, 1998), prepositions are split into categories according to the semantics of the prepositional phrase they may introduce. One of the categories is **spatial**. The prepositions are then further split into modifiers of predicative heads (verbs and predicative nominals) and non-predicative heads (non-predicative nominals). The authors thus emphasize the difference mentioned above. The spatial relations indicated are the following:

- **for predicative heads**
  - position: *The report is <u>on the table</u>.*

  - goal: *The boy is looking towards the beach.*
  - origin: *From this window the views are magnificent.*
  - path: *He ran through the fields.*

- **for non-predicative heads**
  - position: *a brick in the wall*
  - goal: *the train to London*
  - origin: *oranges from Spain*
  - path: *a road through the desert*

We regard all these as instances of the same relations, according to our extension in the definition of locality.

In analyzing spatial relations, Talmy (1985) concentrates on the entities involved in the relation, and suggests different situations based on this analysis. Both closed-class (prepositions) and open-class (verbs) elements may indicate the relative position of the entities involved. Deciding on a relation between objects imposes a referential point. This will cause one (or more) of the entities in the scene to become *Grounds* (referential objects), while one will become the *Figure* (the object that will be located with respect to the other ones). Talmy opposes Fillmore's proposal for **Spatial** relations – Locative, Source, Path, Goal – because they concentrate on direction, rather than the entities involved. For example, all the following examples display Fillmore's Path relation, whereas they would be different in Talmy's case because of the different types of referents (*Ground*):

185 *The ball rolled across the crack.*

186 *The ball rolled past the TV.*

187 *The ball rolled around the lamp.*

In his analysis, Talmy identifies a *small set of primitive station/motion formulas - ones that seem to underlie all more complex characterization of stasis and movement in language.* Talmy refers only to location of entities:

- entity be-located at a point;
- entity directed toward a certain point;

- entity starting at a certain point;
- entity passing through a certain point;
- entity passing/unfolding through a certain location.

Our distinctions are not as fine as the ones Talmy makes in the geometry and relative positions of the entities involved. Further analysis may be done and the relations we propose refined.

4.4.2 Spatiality Case Study

Entities and occurrences can be localized in space. The **Spatiality** relations name the link between an occurrence or an entity, and a region in space. Occurrences can be expressed by clauses or nominal phrases, following *nominalization* derivations. Locality can be expressed by a prepositional or noun phrase, or by an adverb. While it is also possible to identify a location by an occurrence that is unfolding there, the clause expressing this occurrence is a nominal clause, which does not qualify it for a clause-level relation.

Figure 4.5 shows the representation we propose for spatial relations.

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{occurrence/occurrence part/entity} \\ \text{FILLER} & \text{occurrence/part/entity} \end{bmatrix} \\ \text{LOCALITY} & \text{locality indicator} \\ \text{INDICATOR} & \text{indicator} \end{bmatrix}$$

Figure 4.5: The essence of spatial relations

LOCALITY can be any type of locational references covered by the semantic relations: *direction, destination, location from, location to, location at, location through, located, orientation.* Figure 4.6 shows the representation in attribute-value format of the sentence in example (188).

188 $\underline{\textit{He drove}}_{occ}$ *to* $\underline{\textit{Toronto}}_{LocationTo}$.

4.4.3 Definitions

- DIRECTION: 1 is directed towards 2. 2 is not the final point. The final point is not specified.

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{drive} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{LOCATION TO} & \text{Toronto} \\ \text{INDICATOR} & \text{to} \end{bmatrix}$$

Figure 4.6: Attribute-value representation for the LocationTo relation

189 (IC) *Look*$_1$ *inside yourself*$_2$ *for the answer.*

(NP) *inward* $_2$ *look*$_1$

- LocationFrom: 1 starts at 2.

190 (IC) *The capital comes*$_1$ *from* *foreign countries*$_2$.

(NP) *foreign*$_2$ *capital*$_1$

- LocationTo: 2 is the end point of 1.

191 (IC) *I went*$_1$ *home*$_2$.

(NP) *homeward*$_2$ *journey*$_1$

- LocationThrough: 1 occurred through 2. 1 is an occurrence, 2 is a non punctual space.

192 (IC) *We traveled*$_1$ *all over Europe*$_2$.

(NP) *travel*$_1$ *all over Europe*$_2$

- Location: 1 is the location of 2. 2 is an occurrence or an entity, 1 is a point in space (or is considered a point).

193 (IC) *My home is*$_2$ *in* *this town* $_1$.

(NP) *home*$_2$ *town*$_1$ (also spelled *hometown*)

- Located: 1 is located at the point indicated by 2. 1 is an occurrence or an entity, 2 is a punctual space (or is considered punctual).

194 (IC) *The storm started*$_1$ *in* *the desert*$_2$.

(NP) *desert*$_2$ *storm*$_1$

- Orientation: 1 is oriented like 2.

195 (IC) *The tree stood*$_1$ *erect*$_2$*despite the heavy ice.*
(NP) *erect*$_2$ *tree*$_1$

4.5 Conjunctive

4.5.1 General Considerations

The relations in this class describe the conjunction or disjunction of occurrences, entities or attributes. The concepts that interact in such a manner must be of the same type (all of them occurrences, or entities, or attributes).

CONJUNCTION and DISJUNCTION can have instances only on the clause and noun phrase level. These two relations cannot have instances at the intra-clause level, because they are assigned to paratactic constructions, i.e. coordinates, not subordinates. Intra-clause level relations, cases, describe the connections between a superordinate (the verb) and a subordinate (the argument). Cases are assigned to hypotactic constructions.

It can be argued that CONJUNCTION and DISJUNCTION at the noun phrase level have a deeper meaning than the one they have at the clause level. The argument is that the entities at the noun phrase level are more deeply connected than the entities at the clause level. The two entities are in a CONJUNCTION or DISJUNCTION relation because of something they share, whereas at the clause level this is not necessarily the case. We can construct examples that are instances of these two relations at the clause level such that the occurrences therein are unconnected, but this is not normally the case in real texts. Occurrences are connected because the speaker intended to make them so, according to some reason that might or might not be evident. They are sometimes connected because they express causality, temporal ordering, but if the system cannot confidently say which one is the case, it will assign the more generic CONJUNCTION/DISJUNCTION relation. It may also be the case that the occurrences are linked because of a more pragmatic reason (for example, the speaker was reminded of both occurrences at the same time), in which case the assignment of a CONJUNCTION relation is not a weaker assignment but the only one possible. At the NP level, the connection between concepts is more obvious.

4.5.2 Conjunctive Case Study

Conjunctive relations name the link between two occurrences. The two occurrences can be expressed by a full clause, or just a noun phrase. The phenomena that account for the different syntactic forms of concepts connected by CONJUNCTION or DISJUNCTION are *metonymy* and *word formation* phenomena. If the unit under analysis consists only of a noun phrase, as in the example:

196 *I like apples and pears.*

we will consider that since the *like* occurrence applies to both *apples and pears*, it will apply to each of them separately. That results in a conjunction of two *like* occurrences, one whose *Object* are *apples* and the other whose *Object* are *pears.* Why is this a valid point of view? If we have a plural *Object* of an occurrence, we can consider it a conjunction of singular objects. To continue with an example similar to example (196) above, let us consider the sentence:

197 *I like these apples.*

These apples is a plural *Object* of the predicate *like.* It consists of a collection of individual apples. This sentence has a distributive reading, in the sense that we can interpret it in the following manner: each apple in the collection is liked, therefore the *like* occurrences applies to each of them.

Figure 4.7 shows in an attribute-value format the essence of **Conjunctive** relations.

$$\left[\begin{array}{ll} \text{OCCURRENCE1} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{Occurrence1} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \text{occurrence part type} \\ \text{FILLER} & \text{occurrence part} \end{array}\right] \end{array}\right] \\ \text{OCCURRENCE2} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{Occurrence2} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \text{occurrence part type} \\ \text{FILLER} & \text{occurrence part} \end{array}\right] \end{array}\right] \\ \text{INDICATOR} & \text{indicator} \end{array}\right]$$

Figure 4.7: The essence of conjunctive relations

We show an instantiated example of the conjunctive representation in Figure 4.8, for the sentence in example (196), repeated here as example (198):

198 *I like* $\underline{\textit{apples}}_1$ *and* $\underline{\textit{pears}}_2$*.*

The structure is the same as the one that captures the elements of a causal relation. This happens because just like **Causality** relations, **Conjunctive** relations name the link between two occurrences.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{like} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{Object} \\ \text{FILLER} & \text{apples} \end{bmatrix} \end{bmatrix} \\ \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{like} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{Object} \\ \text{FILLER} & \text{pears} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{and} \end{bmatrix}$$

Figure 4.8: Attribute-value representation for the CONJUNCTION relation

4.5.3 Definitions

- CONJUNCTION: both 1 and 2 occur or exist.

199 (CL) *The computer runs applications*$_1$ *and the printer prints documents*$_2$.
(NP) *running*$_1$ *and swimming*$_2$ *(are good for you)*

- DISJUNCTION: either one or both 1 and 2 occur or exist.

200 (CL) *The program may terminate*$_1$ *or it may hang indefinitely*$_2$.
(NP) *painting*$_1$ *or drawing*$_2$

4.6 Participant

4.6.1 General Considerations

These relations describe the interaction between an occurrence and each of its participants. The occurrence can be expressed through a verb, deverbal noun or deverbal adjective. The participant can be an entity or an occurrence.

Participant is one of the most common classes of semantic relations discussed in the literature. It encompasses the relations spun out of grammatical cases: AGENT, OBJECT, INSTRUMENT, etc.

The number and type of semantic relations in this class vary according to the goal pursued. We encounter the relations in this class only at the intra-clause and the noun phrase levels, as they by definition hold between the occurrence and one of its participants. As discussed in Chapter 3, studies of these relations were not concerned with exploring the phenomena that lead to the existence of semantically close clauses and noun phrases that we have identified as *deletion* and *word formation* (nominalization, adjectivalization, etc.)

Macleod et al. (1998) research the mapping between the syntactic arguments of a verb and the modifiers of the nominalized form of the verb. Hull and Gomez (1996) and Gomez (1998a) look at the semantic relations between a verb and its arguments and the ones between the corresponding deverbal noun and its modifiers, to help them distinguish between verbal and non-verbal senses of nominalizations. The interest however is not in exploring the type of transformations that allow for the same relation to be equivalently expressed on the intra-clause and the noun-phrase level.

In a different view of semantic relations, Barker and Dowty (1993) actually search for thematic roles between a noun and its modifier that cannot appear between a verb and one of its arguments. Their general view of the thematic roles is different, and they group all verb-argument roles into two cluster-concepts – *Proto-Agent* and *Proto-Patient*. In order to find roles between a noun and its modifier that are exclusive to noun phrases, they look at *ultra-nominals – nouns which are least plausibly derived from verbs, and those denoting an object.* The nominal thematic roles found are grouped into another two cluster-concepts: *Proto-Part* and *Proto-Whole*. Each is characterized by certain properties, and the position of each of the arguments in the noun phrase is determined by the *nominal argument selection principle* – " in (ultra-nominal) relational nouns, the argument for which the predicate denoted by the noun entails the greatest number of *Proto-Whole* properties will be lexicalized as the object of the preposition *of*, or as the prenominal possessor; the argument having the greatest number of *Proto-Part* entailments will be lexicalized as the head argument".

The first scholar whom we know to have analyzed relations of a semantic nature, Panini, has identified (besides the *karakas* – the verbal relations) non-verbal relations which he grouped under the name *sesa*. Bhartrhari who studied Panini's work in 7th century A.D. further analyzed this group of relations, and affirms that although these relations do not represent karaka relations, they may involve or be preceded by one of the karaka relations. For example, the possessive case is supposed to have been preceded by some action: the possessive expression "king's man", implies that there was some action on the part of the king that has lead to the establishment of a master-servant relation. In expressions like "branch of the tree" and "father's son", the relations like part and whole, and procreator and offspring are "the results of previous actions not mentioned in the sentences, actions in which these objects were accessories. That previous status lingers somewhat in the present status and that is why the present status is looked upon as a kind of *karaka*, though its relation with the action expressed in the sentence is rather remote" (from (Iyer, 1969) cf. (Manjali, 1997)).

Participant relations do not appear at the clause level, because these are relations that involve participants in occurrences. There are, however, instances in which the participant is expressed by an occurrence, as is the case for the *Object* of predicates like *know, say*, etc.

4.6.2 Participant Case Study

Participant relations name the connection between an occurrence and the entities involved. A participant in an occurrence can be expressed using a nominal clause that specifies a generic concept. For example:

201 *The one who was painting the walls got sick.*

The clause *who was painting the walls* serves to restrict the general concept *the one* to a particular individual. The participant relation will hold between *the one* and the predicate *got sick*, and the clause will be a postmodifier for the noun *one.*

Certain verbs can take an occurrence as an argument:

202 *I know what you did last summer.*

One of the arguments of a **Participant** relation must be an occurrence. It must surface somehow or be implied in the syntactic expression of an occurrence-participant pair. The phenomena that account for the different surface realizations of an occurrence are *word formation* (*nominalizations*, *adjectivalizations*), *deletion*, and *metonymy.* The participant may also surface in different syntactic structures. *Adjectivalization* is a phenomenon that allows for a concept to surface as a modifier in a noun-phrase.

In Figure 4.9 we show the structure that captures the essence of **Participant** relations. An instantiated example is presented in Figure 4.10.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{Occurrence} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{occurrence part type} \\ \text{FILLER} & \text{occurrence part} \end{bmatrix} \end{bmatrix} \\ \text{PARTICIPANT-ROLE} & \text{Participant} \\ \text{INDICATOR} & \text{indicator} \end{bmatrix}$$

Figure 4.9: The essence of participant relations

PARTICIPANT-ROLE will be one of: *Agent, Co-Agent, Beneficiary, Exclusion, Stative, Property, Possessor, Possession, Object, Object-Property, Instrument, Recipient, Product*, which describe the role of the participant in the occurrence.

203 *I eat supper*$_{occ}$ *with* *my family*$_{CoAgent}$.

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{eat} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \text{Object} \\ \text{FILLER} & \text{supper} \end{array}\right] \end{array}\right] \\ \text{CO-AGENT} & \text{family} \\ \text{INDICATOR} & \text{with} \end{array}\right]$$

Figure 4.10: Attribute-value structure for the CO-AGENT relation

PART and WHOLE are exceptions, in that the above structure does not apply. The reason is that they occur only at the noun phrase level, and they link two entities, one of which represents the whole and the other the part. They will have the structure shown in Figure 4.11.

Figure 4.12 shows a sample representation for the noun phrase:

204 *board*$_{whole}$ *member*$_{part}$

4.6.3 Definitions

- CO-AGENT(accompaniment): 1 is accompanied by 2. 2 is a co-agent.

  205 (IC) *We eat supper*$_1$ *with my family*$_2$.
  (NP) *supper*$_1$ *with my family*$_2$

- AGENT: 1 performs 2.

  206 (IC) *The students*$_1$ *protested*$_2$ *against tuition fee increase.*
  (NP) *student*$_1$ *protest*$_2$

$$\left[\begin{array}{ll} \text{ENTITY1} & \text{WHOLE entity} \\ \text{ENTITY2} & \text{PART entity} \\ \text{INDICATOR} & \text{indicator} \end{array}\right]$$

Figure 4.11: The essence of WHOLE and PART relations

$$\begin{bmatrix} \text{WHOLE} & \text{board} \\ \text{PART} & \text{member} \\ \text{INDICATOR} & - \end{bmatrix}$$

Figure 4.12: Attribute-value representation for the PART relation

- BENEFICIARY: 1 benefits from 2.

207 (IC) *The price discount applies*$_2$ *only for students*$_1$.
(NP) *student*$_1$ *discount*$_2$

- EXCLUSION: 2 is excluded from 1, or 1 replaces 2.

208 (IC) *We cooked rice*$_1$ *instead of potatoes*$_2$.
(NP) *rice*$_1$ *instead of potatoes*$_2$

- STATIVE: 1 is in a state of 2.

209 (IC) *The dog*$_1$ *is sleeping*$_2$.
(NP) *sleeping*$_2$ *dog*$_1$

- PROPERTY: 1 has the property 2.

210 (IC) *The dog*$_1$ *is brown*$_2$.
(NP) *brown*$_2$ *dog*$_1$

- POSSESSOR: 1 has 2.

211 (IC) *The man has*$_1$ *a long beard*$_2$.
(NP) *bearded*$_2$ *man*$_1$

- POSSESSION: 2 has 1.

212 (IC) *The nation has*$_2$ *a big debt*$_1$.
(NP) *national*$_2$ *debt*$_1$

- INSTRUMENT: 1 uses 2.

213 (IC) *The system administrator notified*$_1$ *the users via e-mail*$_2$.
(NP) *e-mail*$_2$ *notification*$_1$

- OBJECT: 1 is acted upon by 2.

214 (IC) *They repair*$_2$ *engines*$_1$.
(NP) *engine*$_1$ *repair*$_2$

The *Object* in an occurrence can also be a clause, as in the example:

215 *They repair*$_2$ *what other people break*$_1$.

Although this is a relation between two clauses, it is not a clause level relation, because the *Object* clause is an element of the occurrence, and not merely a related occurrence.

- Object-Property: 1 was acted upon by 2.

216 (IC) *They repaired*$_2$ *the engine*$_1$.
(NP) *repaired*$_2$ *engine*$_1$

The difference between Object and Object-Property may seem vague. In the Object-Property relation, the entity involved is specific. At the intra-clause level this is signaled by using specific determiners (*the, that*, etc.), and at the noun-phrase level by identifying the *Object* through an action which it underwent. The *Object* in an Object relation is a generic entity both at the intra-clause and at the noun-phrase level.

- Recipient: 2 receives the object of 1.

217 (IC) *We wrote*$_1$ *Smilla a reference letter to prospective employers*$_2$.

- Part: 1 is part of 2.

218 (NP) *the funnel*$_1$ *of the ship*$_2$

- Product: 1 produces 2.

219 (IC) *The factory builds*$_1$ *cars*$_2$.
(NP) *car*$_2$ *factory*$_1$

- Whole: 2 is part of 1.

220 (NP) *daisy*$_2$ *chain*$_1$

4.7 Quality

4.7.1 General Considerations

This class groups relations that describe different attributes (such as measure, type, order) associated with concepts One of the relations in this group, MANNER, describes the property of an unfolding action. One of the elements it links must be an action expressed by a verb in some form. The property surfaces as an adverb or adjective. All the other relations in this class are best described as holding between two of the arguments of the main verb, rather than between the verb and its arguments. The verb is there for purely syntactic reasons, or to explain better the relation between other entities, but does not play an important role itself.

The grouping of relations into the **Quality** class belongs to us, and as opposed to the more coherent and commonly agreed upon classes – Causal, Temporal, Spatial, Participant.

Whereas there is no consensus in the field on a list of semantic relations for either syntactic level, in most works, the first 5 classes of relations are largely agreed upon.

We find however a number of these relations mentioned in literature. Quirk et al. (1985, p. 557) provides examples of MANNER relations introduced by various adjuncts:

221 *They began arguing loudly.*
You should write as I tell you to.

Thompson and Longacre (1985) also show various expressions of the MANNER relation through various forms in different languages (see Section 3.4.3.2).

PROPERTY is an obvious relation. We all associate and also use properties to identify objects: *blue book*, *hard project*, etc. It is one of the basic relations associated with the verb *be*, which can be deleted to obtain a more compact phrase (Levi, 1978).

TYPE is the name we gave the hyponym-hypernym relation (IS-A) – *maple leaf.*

MATERIAL is also a recognized relation. Vanderwende (1994), for example, includes a relation called *Made of what* in her list, which would correspond to our MATERIAL relation.

MATERIAL is also included by Leonard (1984), so is EQUATIVE.

4.7.2 Quality Case Study

Quality relations hold between an occurrence and one of its arguments. The argument shows aspects of the occurrence – manner, type, measure, etc. The occurrence must surface in the syntactic expression. The phenomena identified in the verbalization of concepts connected by this type of relations are *metonymy* and *word formation* processes.

We present the essence of **Quality** relations in Figure 4.13.

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{Occurrence} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \text{occurrence part type} \\ \text{FILLER} & \text{occurrence part} \end{array}\right] \end{array}\right] \\ \text{QUALITY-REL} & \text{Quality} \\ \text{INDICATOR} & \text{indicator} \end{array}\right]$$

Figure 4.13: The essence of quality relations

QUALITY-REL is one of the relations in the **Quality** class.

Figure 4.14 shows an instantiation of the attribute-value representation for the example:

222 $\underline{\textit{a hundred-dollar}}_{measure}$ $\underline{\textit{book}}_{occ}$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{book} \end{array}\right] \end{array}\right] \\ \text{MEASURE} & \text{hundred-dollar} \\ \text{INDICATOR} & - \end{array}\right]$$

Figure 4.14: Attribute-value structure for the MEASURE relation

The structure in Figure 4.13 does not apply for the EQUATIVE and TYPE relations. The reason is that they connect two entities, and not an occurrence and one of its attributes. EQUATIVE and TYPE occur only at the noun phrase level, and will have the structure presented in Figure 4.15:

$$\left[\begin{array}{ll} \text{ENTITY1} & \text{Entity 1} \\ \text{ENTITY2} & \text{Entity 2} \\ \text{INDICATOR} & \text{indicator} \end{array}\right]$$

Figure 4.15: The essence of EQUATIVE and TYPE relations

The attribute-value structure for the sentence in example (223) is shown in Figure 4.16.

223 *cumulus*$_{type}$ *cloud*$_{entity}$

$$\begin{bmatrix} \text{ENTITY} & \text{cloud} \\ \text{TYPE} & \text{cumulus} \\ \text{INDICATOR} & - \end{bmatrix}$$

Figure 4.16: Attribute-value representation for the Type relation

4.7.3 Definitions

- Content(physical content):1 contains 2.

224 (IC) *He filled the bottle*$_1$ *with milk*$_2$.

(NP) *milk*$_2$ *bottle*$_1$

- Topic(abstract content): 1 is concerned with 2.

225 (IC) *John produced a documentary*$_1$ *about volcanoes*$_2$.

(NP) *volcano*$_2$ *documentary*$_1$

- Container: 1 is contained in 2.

226 (IC) *He poured milk*$_1$ *into the bottle*$_2$.

(NP) *bottle*$_2$ *of milk*$_1$

- Manner: 1 occurs in the way indicated by 2.

227 (CL) *You should write*$_1$ *as I tell you*$_2$.

(IC) *You write*$_1$ *with style*$_2$.

(NP) *stylish*$_2$ *writing*$_1$

- Material: 1 is made of 2.

228 (IC) *We build houses*$_1$ *with bricks*$_2$.

(NP) *brick*$_2$ *houses*$_1$

- Measure: 2 is a measure of 1

229 (IC) *The car cost*$_1$ *five hundred dollars*$_2$.

(NP) *five-hundred dollar*$_2$ *car*$_1$

- ORDER: 1 is before 2 in physical space.

230 (IC) *He filed the Baker file*$_1$ *before* *the Abel file*$_2$.
(NP) *A files*$_1$ *before* *B files*$_2$

- EQUATIVE: 1 is also 2.

231 (NP) *composer*$_2$-*arranger*$_1$

- TYPE: 1 is a type of 2.

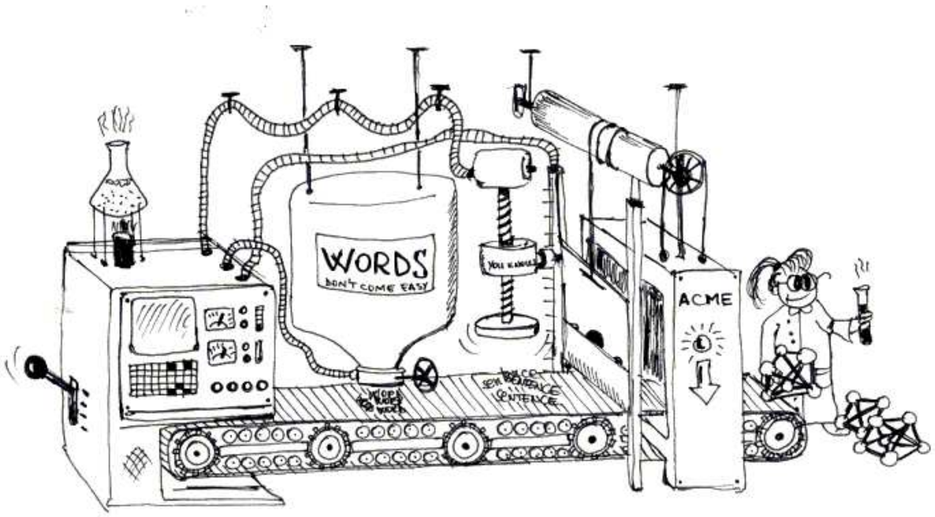
232 (NP) *oak*$_1$ *tree*$_2$

4.8 Conclusions

We have shown in this section how we can use the phenomena identified in Chapter 3 to show that semantic relations are the same across syntactic levels. We have grouped semantic relations according to their meaning into six classes: causal, temporal, spatial, participant and quality. For each class we have identified various possibilities of connecting two concepts, and we have shown surface expressions for the same concepts (as much as possible) on different syntactic levels to demonstrate that the same semantic relation can appear on different syntactic levels. Also, for each class we have proposed a structure that describes the essence of the semantic relations between concepts in that particular class.

Chapter 5

Testing a List of Semantic Relations That Spans Three Syntactic Levels



5.1 Introduction

In the preceding chapters we showed that it is valid to view semantic relations as not depending on syntactic levels. We will show, using a system that acquires knowledge from texts, that this view has a positive practical impact: it improves the knowledge acquisition process.

The theoretical discussion presented in Chapters 3 and 4 show principles and phenomena that allow us to explain why semantic relations can have instances at various syntactic levels. Based on these ideas, we show in this chapter how to combine lists of semantic relations that pertain to different syntactic levels into one list that covers all these levels. We will then test the combined list in the context of knowledge acquisition.

We work with TANKA – a text analysis and knowledge acquisition system – which separately analyzes each of these levels: clause, intra-clause and noun-phrase. Each level uses a list of semantic relations specifically designed to capture the interaction between syntactic units specific to this level. TANKA and the lists of semantic relations it uses are described in (Delisle, 1994) and (Barker, 1998).

The experiment we present in this chapter is meant to show that having a combined list of relations across several syntactic levels, improves the process of knowledge acquisition. Since there are no syntax-based distinctions, every bit of knowledge acquired is placed in the same "pool" of knowledge. In analyzing the next bit, everything that was processed before is now available, and intuitively, this should help the knowledge acquisition process. The results of this experiment show that this intuition is correct. A system that uses one list of relations and treats the text uniformly, without distinguishing between syntactic levels, learns faster than a system that analyzes syntactic levels separately, and keeps the knowledge gathered separated according to the syntactic level on which it was found.

We should underline the fact that for the purpose of our experiment, only the process of knowledge acquisition is evaluated, and not the result obtained. Processing the text with three different lists, or just one list, has no effect on the number of pairs of concepts extracted, and the number of pairs added to the knowledge base. Without reference resolution, which would allow us to collapse into one node all instances of a concept, we do not achieve the level of compaction in the knowledge representation that would allow us to claim an improvement.

The system we use for this experiment, TANKA, has two main components: a syntactic analyzer (DIPETT) and a semantic analyzer (HAIKU). For the purpose of this experiment, HAIKU must be modified so that it uses only one list of relations for pairs of entities extracted from any of the syntactic levels for which DIPETT provides information. In Section 5.2 we show how we obtain a list of relations that spans three syntactic levels, by combining three separate lists, one for each of these levels. We present the three lists with which we start, and then show the modifications introduced as a consequence of the unified view we propose.

In Section 5.3 we show the modifications we brought to HAIKU, the semantic analysis module and then we show the experiment performed. We discuss in Section 5.4 the results obtained and their significance.

5.2 A Combined List of Relations

The list of relations that we are putting to the test brings together three lists of relations described in (Barker, 1998).

The first indication that it is necessary to combine the three lists of relations was the fact that at different syntactic levels there were relations with the same name. A closer look into this matter has lead us to uncover several phenomena in verbalization that support the view that semantic relations can be the same, even if the entities they connect have different embodiments in syntactic structures.

The unification process started with manual analysis in which semantic relations from the three different lists were analyzed and compared with each other to discover overlaps and commonalities. The theoretical analysis helped identify the requirements of each relation, with respect to the nature of their arguments – that is, the concepts they connect.

The result of this analysis is a combined list of semantic relations that covers all the relations from the three initial lists. We present it in Table 5.1.

In the remainder of this section we will show how we have mapped the three lists we started with into one that spans the clause, intra-clause and noun-phrase level. We have looked at each relation and tried first to see if we can find an instance of it on other syntactic levels. If that was possible, we have tried to find a counterpart for it in the other lists. This has lead us to rename or split some of the existing relations, to be able to align the three lists, and map them into one.

5.2.1 Causality

Table 5.2 shows the correspondence between the causality relations in the three original lists, and the combined list[1].

[1] In each of the tables presented in Section 5.2 the notational convention is as follows. The row corresponding to each syntactic level contains the relations in the original list for that level. The columns correspond to all relations possible for that class, which appear in the combined list. A relation name in a cell at position (i, j) shows a relation from the original list corresponding to level i, and that particular comment. If the cell at position (i, j) in the table is empty that means in the original lists there was no relation corresponding to syntactic level i and column j, but we have found an instance of the relation. If the cell contains the $\nexists$ symbol, it means that neither the original lists, nor the combined one has a relation to cover that particular situation.

A relation on a certain level can span several columns. That relation will be split into the relations corresponding to the columns it spans from the combined list. This is the case for the NMR Time, for example, which spans 7 columns in Table 5.3, which shows the **Temporal** relations. It will be split into 7 more specific relations

Relation	Abbr.	Example	Paraphrase
CAUSALITY			
Cause	cs	flu virus	H makes M occur or exist, H - necess. and suff.
Effect	eff	exam anxiety	M makes H occur or exist, M - necess. and suff.
Purpose	prp	concert hall	H is for V-ing M, M - not necess. occurs or exists
Entailment	ent		H makes M occur or exist, H - not known to exist
Detraction	detr	headache pill	H opposes M, H - not suff. to prevent M
Prevention	prev		H opposes M, H - suff. to prevent M
TEMPORALITY			
Co-occurrence	cooc		H and M occur or exist at the same time
Frequency	freq	daily exercise	H occurs every time M occurs
Precedence	prec		H (begins to) occurs or exists before M
TimeAt	tat	morning exercise	H occurs when M occurs
TimeFrom	tfr		H began to occur when M became true
TimeThrough	tthr	six-hour meeting	H existed while M existed, M - interval of time
TimeTo	tto		H existed until M started to exist
SPATIAL			
Direction	dir	outgoing mail	H is directed towards M, M is not the final point
Location	loc	home town	H is the location of M
LocationAt	lat	desert storm	H is located at M
LocationFrom	lfr	foreign capital	H originates at M
LocationTo	lto		the destination of H is M
LocationThrough	lthr		H occurred through M (M is a space)
CONJUNCTIVE			
Conjunction	conj		both H and M exist
Disjunction	disj		either one or both H and M exist
PARTICIPANT			
Co-Agent	acc		H is accompanied by M (co-agent)
Agent	ag	student protest	M performs H, M - animate or natural phen.
Beneficiary	ben	student discount	M benefits from H
Exclusion	excl		M is excluded from H, or H replaces M
Instrument	inst	laser printer	H uses M
Object	obj	metal separator	M is acted upon by H
Object-Property	obj prop	sunken ship	H underwent M
Part	part	printer tray	H is part of M
Possessor	posr	national debt	M has H
Property	prop	blue book	H is M
Product	prod	plum tree	H produces M
Source	src	olive oil	M is the source of H
Stative	st	sleeping dog	H is in a state of M
Whole	whl	daisy chain	M is part of H
QUALITY			
Container	cntr	film music	M contains H
Content	cont	apple cake	M is contained in H
Equative	eq	player coach	H is also M
Manner	man	stylish writing	H occurs in the way indicated by M
Material	mat	brick house	H is made of M
Measure	meas	expensive book	M is a measure of H
Order	ord		H is before M in physical space
Topic	top	weather report	H is concerned with M
Type	type	oak tree	M is a type of H

Table 5.1: Table of semantic relations with examples

Level	**Relations**						
Clause	Causation		Detraction	Enablement	Entailment	Prevention	
Intra-clause	Cause	Effect	Opposition				Purpose
Noun-phrase	Cause	Result		∄			Purpose
Combined	Cause	Effect	Detraction	Enablement	Entailment	Prevention	Purpose

Table 5.2: Combining **Causality** relations

- Cause and Effect: Following the model for the intra-clause and noun-phrase definitions of the Cause and Effect relations, we have split the Causation clause-level relation into two relations, to parallel the other two syntactic levels.

  Both for cases and noun-modifier relations (NMRs), it is the argument that names the relation: if the argument is the Cause, the relation between the argument and the head will be Cause, and the same for Effect.

  Cause was originally defined as an NMR:

  "*modifier* causes *compound*"
  and Result:

  "*modifier* is a result of *compound*"

  At the intra-clause level, if the argument is the state or event that causes the occurrence described by the verb to take place, the relation between them will be Cause. If the argument describes the state that results from the occurrence described by the verb, the relation will be Effect.

  We will split Causation as follows. If the subordinate clause describes the *causing* occurrence, and the main clause describes the *resulting* one, the relation will be Cause. If the situation is reversed, the relation will be Effect.

  For example:

  233 *The program issued a command, so the file printed.*

  234 *The file printed because the program issued a command.*

  In both sentences, the *Cause* clause is *the program issued a command*, and the *Effect* is *the file printed.* In example (233) the *Effect* clause is the subordinate, in example (234), the *Cause* clause. Following the intra-clause and noun-phrase level conventions, the Effect relation describes the connection between the clauses in sentence (233), and Cause the link between the clauses in sentence (234).

- Detraction and Opposition: The clause-level relation (CLR) Detraction and the case Opposition can be mapped onto each other. Their original definitions are as follows.

  > Detraction: The Detraction relationship represents the situation when E_1 detracts from or opposes E_2 but is insufficient to prevent E_2 from occurring or existing.
  >
  > **235** *Although* $\underline{\textit{the server was very busy}}_{E_1}$, $\underline{\textit{the program ran}}_{E_2}$.
  >
  > Opposition: The Opposition case represents an entity that contrasts with or opposes the event, but is insufficient to prevent it from happening.
  >
  > **236** *Despite* $\underline{\textit{my warning}}_{OPP}$, *they persisted.*
  >
  > (Barker, 1998)

  The entity in the definition of Opposition can stand for a whole occurrence (through *metonymy*) or *nominalization*.

  We have also found instances of the Detraction relation at the noun-phrase level: *persistence despite warnings.* We have adopted the name Detraction to cover this relation on all levels.

- Enablement: this relation describes the interaction between two occurrences, in which the existence of one is necessary but not sufficient for the existence of the other. We have found instances of this relation at the intra-clause level, but not at the noun-phrase level, as shown in example (165), repeated here as example (237):

  237 (CL) *The printer can print if the paper tray is full.*
  (IC) *The printer can print from a full paper tray.*

  One explanation of the fact that we have found no instance of this relation at the noun-phrase level may be the fact that we need conditionals to express this relation. Conditional expressions need a modal verb in English (for example *can*), and therefore cannot surface at the noun-phrase level.

- Entailment: the relation describes the link between two occurrences, in which the existence of one is necessary and sufficient for the existence of the other. The relation has instances at all levels:

238 (CL) *If you eat sweets before dinner, you spoil your appetite.*
(IC) *Sweets before dinner spoil your appetite.*
(NP) *lack of appetite because of sweets*

- Prevention: the existence of one occurrence is enough and sufficient to prevent the other occurrence from existing. We have found instances of this relation on other levels than CL, through phenomena of metonymy and nominalizations.

5.2.2 Temporality

Some of the relations that existed at one syntactic level have been split according to the type of the time intervals (punctual, finite, open-ended) represented by the interacting occurrences. In some cases, the name of the original relation was assigned to one of the subsets, because it described very well the more constrained relation.

- Time (NMR). At the noun-phrase level, there was only one temporal relation, generically named Time. We have found noun-phrase expressions of all the temporal relations from the clause and intra-clause level.

  - TimeAt: *evening news*;
  - TimeTo: *party till dawn*;
  - TimeThrough: *winter semester*;
  - TimeFrom: *party since yesterday*;
  - Frequency: *daily news*;
  - Precedence: *coffee after 5 o'clock*;
  - Co-occurrence: *singing while showering*;

  We have therefore decided to split the original Time NMR into the seven fine-grained temporal relations we have found at the other two levels.

- Co-occurrence (CLR). At the clause level there were originally the following two relations: Co-occurrence and Precedence. We have found instances of other temporal relations at the clause level:

  - TimeAt: *He went there when they called him.*
  - TimeTo: *We will travel until we ran out of money.*
  - TimeThrough: *He sat patiently while the soprano sang an aria.*
  - TimeFrom: *I have been waiting to meet you since I saw you on TV.*

  - Frequency: *We play volleyball every time he visits.*
  - Co-occurrence: *He paints while his wife plays the piano.*

  The distinctions between the different temporal relations come from the type of time interval they cover. For TimeAt, the temporal reference is considered punctual, as it is for TimeTo, TimeFrom and Frequency. The difference between these four relations is the relative position of the occurrence and the temporal reference. TimeThrough and Co-occurrence involve time intervals. For TimeThrough the intervals are bounded, for Co-occurrence they are not.

- Precedence(CLR). There was no Precedence relation at the intra-clause level. Precedence is defined at the clause level as showing the relative position of two occurrences, one precedes the other on the time axis. We have found instances of this relation at the noun-phrase level, as we have shown earlier, and also at the intra-clause level:

  - Precedence: *I never drink coffee after 5 o'clock.*

A compact view of the result of the combination process is presented in Table 5.3.

Level	**Relations**						
Clause	Co-occurrence						Precedence
Intra-clause		TimeAt	TimeFrom	Frequency	TimeThrough	TimeTo	
Noun-phrase	Time						
Combined	Co-occr	TimeAt	TimeFrom	Frequency	TimeThrough	TimeTo	Precedence

Table 5.3: Combining **Temporal** relations

5.2.3 Spatiality

Some of the **Spatial** relations on the noun phrase level have been renamed, and a few introduced.

- Destination (NMR): the relation describes the interaction between a noun and its modifier, the modifier indicates the destination of the concept expressed through the noun. There is no distinction between the situation when the modifier indicates an end point, and when it indicates a general direction. We find this distinction at the intra-clause level, and we split the Destination accordingly into:

- Direction: the modifier expresses a general direction of movement.

  239 *northward journey*

- LocationTo: the modifier expresses the end point of the movement.

  240 *homeward journey*

- Source (NMR): this relation also seems to cover two different aspects: spatial and origin. Each of the two aspects of the Source relation has its own correspondent at the intra-clause level. We will use these two relations to differentiate the two situations covered by the Source relation:

  - LocationFrom: covers the spatial aspect.

    241 *outside air*

  - Source: covers the origin aspect. This restricted version of the original Source NMR belongs to the **Quality** class of relations.

    242 *olive oil*

- LocationThrough (Case): we have found instances of this relation at the noun-phrase level.

  243 *world-wide travel*

- Orientation (Case): was introduced at the NP level.

  244 *tilted tower*

Table 5.4 shows in a more concise way the combination of **Spatiality** relation across syntactic levels.

Level	**Relations**						
Clause	∄	∄	∄	∄	∄	∄	∄
Intra-clause	LocationThr	LocationAt		LocationFrom	Direction	LocationTo	Orientation
Noun-Phrase		Location	Located	Source	Destination		
Combined	LocationThr	Location	Located	LocationFrom	Direction	LocationTo	Orientation

Table 5.4: Combining **Spatiality** relations

5.2.4 Conjunctive

Both Conjunction and Disjunction have been introduced for the NP level.

245 *running and swimming [are good for you]*

246 *[he can make either] a painting or a drawing [of you.]*

These two relations can have instances only at the clause and noun-phrase level, because these levels that allow paratactic constructions. Cases are by their nature hypotatic relations – they connect a superordinate with a subordinate.

Table 5.5 shows the combined list of relations for the **Conjunctive** class.

Level	**Relations**	
Clause	Conjunction	Disjunction
Intra-clause	∄	∄
Noun-phrase		
Combined	Conjunction	Disjunction

Table 5.5: Combining **Conjunctive** relations

5.2.5 Participant

Participant relations appear only at the intra-clause and noun-phrase levels. We have explained this in Section 4.6. Of these relations, Part and Whole appear only at the NP level. We have modified the original list of NMRs and cases as follows, in order to combine them.

- Accompaniment was added to the list of NMRs, and renamed Co-Agent. It describes the participant in an occurrence that performs the specified activity together with the *Agent*:

  247 (IC) *I traveled to Switzerland with my friend.*
  (NP) *travel with my friend*

- Exclusion was added to the list of NMRs. Exclusion is used to pinpoint a participant that is (explicitly) excluded from an occurrence. The activity captured in the occurrence is the factor that imposes the distinction between participants, but the Exclusion relation is more of a set-member relation. It holds between the excluded participant and a set of (or just one) other (accepted) participants. Such a relation can be expressed at the NP level as well.

248 (IC) *Everybody slept except* $\underline{\textit{Jeff}}_{EXCL}$
(NP) *everybody except Jeff [slept]*

- Experiencer (Case): This relation represents the (non-active) participant that is experiencing a state or sensation. We have observed that the more general Experiencer case corresponds to four more specific NMRs:
  - Stative: The experiencer is in a certain state.

    249 *The log is floating. ↔floating log*
  - Property: The experiencer has a certain property.

    250 *The girl is pretty. ↔pretty girl*
  - Possession: The experiencer has (owns) something.

    251 *The company has a car. ↔company car*
  - Possession: The experiencer has (owns) something. The experiencer is the head of the noun-phrase.

    252 *The man has a beard. ↔bearded man*

- Possession (NMR): we have introduced this subset of the Property NMR, to describe the relation between a participant and an acquired property, about which we can say that the participant possesses it. An example on the intra-clause and noun-phrase level is given in (252).

- Product was added to the list of cases, as a subset of the more general Object relation. It describes the entity that was obtained as a result of some activity/event occurrence.

- Object-Property was introduced as a special instance of the Object case and the Property NMR. Similarly to Possession, it describes the relation between a participant and an acquired property. The difference is that we cannot say that the participant possesses this property, as shown in the examples:

  253 *The treasure vanished. ↔vanished treasure*

- Instrument was split into Instrument and Co-Instrument. We have observed while analyzing verb definitions from *LDOCE* that one could distinguish between an actual instrument used in an occurrence, and something that is part of an instrument,

Level	**Relations**					
Intra-clause	Accompaniment	Agent	Beneficiary	Instrument		Recipient
Noun-phrase		Agent	Beneficiary	Instrumental		∄
Combined	Co-Agent	Agent	Beneficiary	Instrument	Co-Instrument	Recipient

Level	**Relations**						
Intra-clause	Experiencer				Object		
Noun-phrase	Stative	Possessor	Property			Object	Product
Combined	Stative	Possessor	Possession	Property	Object-Property	Object	Product

Level	**Relations**		
Intra-clause	Exclusion	∄	∄
Noun-phrase			
Combined	Exclusion	Part	Whole

Table 5.6: Combining **Participant** relations

but not an instrument itself. As an example, consider the distinction between the roles of *ink* and *pen* in a writing occurrence:

254 *I wrote the letter with ink.*
I wrote the letter with a pen.

Ink cannot be regarded as an instrument, but as part of one, while *pen* can.

- Part and Whole: we have introduced these relations at the NP level. The discussion presented in Section 4.6 shows that these relations can only appear at the NP level.

255 *board member (*Part*)*

256 *ultraviolet light (*Whole*)*

Table 5.6 gives a more concise view of the operations we have performed on the relations in the three original lists for alignment purposes.

5.2.6 Quality

The following modifications have been brought to the original lists of relations.

- Content (case): one of the elements of the relation is an entity that expresses the topic of a communication or consideration event, and also the physical content of a container. At the NP level, there are two different relations, one for each of the two aspects of Content. There is also a relation that covers the involvement of the container in an occurrence. We have decided to split the Content relation into the following subsets:

- Content: covers physical content.

  257 (IC) *I filled the bottle with* $\underline{\textit{milk}}_{CONT}$. ↔(NP) *milk bottle*

- Topic: covers abstract content.

  258 (IC) *He produced a documentary about* $\underline{\textit{volcanoes}}_{TOP}$. ↔(NP) *volcano documentary*

- Container: the relation links the container with the occurrence.

  259 (IC) *They added music to* $\underline{\textit{the film}}_{CNTR}$. ↔(NP) *film music*

- Manner (case): The manner in which an occurrence should unfold can be expressed through an occurrence, an adverb or adjective. As a consequence, we have introduced this relation at the clause and noun-phrase levels:

  260 (CL) *He draws like* $\underline{\textit{his teacher told him}}_{MAN}$.
  (IC) *He draws* $\underline{\textit{beautifully}}_{MAN}$.
  (NP) $\underline{\textit{beautiful}}_{MAN}$ *drawing*

  At the NP level, this relation was covered by Property. We distinguish Manner as a special case of Propertythat applies to occurrences, as opposed to entities and their attributes. The head noun in a noun phrase tagged with the Manner relation expresses an occurrence through a verb-derived nominal.

- Measure (case): we distinguish another subset of the Property relation, which parallels the Measure case. In this situation, the modifier expresses the measure of an occurrence associated with an entity:

  261 (IC) *I bought the car for* $\underline{\textit{five hundred dollars}}_{MEAS}$.
  ↔(NP) $\underline{\textit{five-hundred dollar}}_{MEAS}$ *car*

- Order (case): we have introduced the Order relation at the NP level, to describe the relative position of two entities in a sequence:

  262 *I have filed the A files before the* $\underline{\textit{B files}}_{ORD}$.
  $\underline{\textit{A files}}$ before $\underline{\textit{B files}}_{ORD}$

- Type (NMR): we have introduced this relation to cover the situation when the modifier in a noun phrase represents an entity which is a particular instance of the more general entity represented by the head noun:

Level	**Relations**				
Clause	∄	∄	∄		∄
Intra-clause	Content			Manner	Measure
Noun-phrase	Content	Topic	Container	[Property]	
Combined	Content	Topic	Container		Measure
Level	**Relations**				
Clause	∄	∄	∄	∄	∄
Intra-clause	Material	Order	∄	∄	Source
Noun-phrase	Material		Equative		Source
Combined	Material	Order	Equative	Type	Source

Table 5.7: Combining **Quality** relations

263 *oak tree, cumulus cloud*

We present the original lists and the combined list of **Quality** relations in Table 5.7.

5.3 Testing the Combined List

In order to test the combined list of relations, the semantic analysis module of TANKA, HAIKU, must be modified to accommodate a uniform treatment of semantic relations and syntactic structures.

Once the new HAIKU is operational, it is plugged into TANKA. For comparison purposes, the experiment will be conducted following the procedure described in (Barker et al., 1997b): we will have two experimenters that assign semantic relations together, the system will conduct the same type of dialog with the users (presented in Section 5.3.1.3.2), we collect the same statistics (presented in Section 5.3.1.3.3), and we use the same input data – a semi-technical text on meteorological phenomena (Larrick, 1961).

We will compare the learning curve of the old and the new system - a steeper curve means faster learning. We expect the system to show a continuous increase in the number of examples it tags correctly with semantic relations, and a flattening of the curve that shows the amount of interaction with the user.

5.3.1 The New HAIKU

Using a combined list of relations over three different syntactic levels imposes the constraint of treating pairs of syntactic structures uniformly, regardless of the syntactic level they were extracted from.

The original version of HAIKU had three different modules, to treat separately each syntactic level.

The new HAIKU will take the input parse trees, and will process them following the algorithm described in Figure 5.1.

```
for each parse tree PT
 {
   postprocess PT;
   extract all pairs of syntactic units;
   analyze each pair;
  }
```

Figure 5.1: New HAIKU algorithm

5.3.1.1 Parse Tree Post-Processing

Parse tree post-processing consists of two phases.

1. **Pronoun resolution**. In the original test, pronoun resolution was performed semi-automatically. Since this process was not modified as a result of the present research, we have bypassed this phase. Pronouns and their referents are extracted from a manually built database, and the system will automatically substitute the corresponding structures in the parse trees.

2. **Conjoined phrases**. If the input parse tree contains conjoined verbs or verb phrases, the system will automatically build separate parse trees for each member of the conjunction/disjunction. Each parse tree will be analyzed separately, and independent from the other ones. The processing of conjoined phrases follows closely the algorithm from the original HAIKU.

5.3.1.2 Extracting Pairs of Syntactic Units

The new HAIKU system diverges from the original one at the following two steps of the algorithm: extracting pairs of syntactic units and analyzing them semantically. The new HAIKU uses the parse trees obtained after the preprocessing stage, and extracts and analyzes pairs of syntactic units in a process that does not discriminate based on the syntactic level.

Figure 5.2 shows the algorithm for extracting pairs of syntactic structures from the parse tree.

We elaborate each processing step.

```
extract all syntactic units from the parse tree
for each syntactic unit S
{
  assign S a unique identifier;
  extract information from each simple structure;
  add the information to a database indexed on word forms;
  build a simplified tree by replacing syntactic units
    with their identifiers;
  traverse the simplified tree and each structure to extract pairs;
}
```

Figure 5.2: Algorithm for extracting pairs of syntactic structures

Step 1 **Extract each syntactic unit from the parse tree**. The system will search for specific structures in the parse tree, using List 1.

List 1 – Syntactic units

[adj,	adjectives
n, proper_noun,	nouns (common, proper)
advs, adv_clause, simple_adv_clause, pp_adv,	adverbial modifiers (simple adverbs, adverbial clauses, adverbial phrases)
entity,	covers anything that can be conceived of as an entity
predicate,	head of a verb phrase
statement,	clause
simple_sentence, complex_sentence,	sentence (simple or complex)
subord_clauses, head_main_clause, next_main_clause]	types of clauses (subordinate, ,main or coordinate)

We have given this list to the system. It consists of non-terminals from the syntactic analyzer's grammar that describe structures of interest to semantic analysis, on all levels of the parse tree[2].

The list is ordered so that the structures that are embedded (in other structures from the same list) are processed first.

For each syntactic unit the system will extract the following information.

- **Head word**. For each structure, extract the head word. The head word is an open-class word. It is the uninflected form of a word in the sentence. Ideally, every open-class word should be lemmatized. The reason is that we

[2] For a comprehensive list of nonterminals in DIPETT's grammar refer to (Delisle, 1994)

want a family of words together, so that information about the relations in which one member of the family participates is available to any member. For example, *parent* and *parental* should be in the same family, and the same is true for *refuse* and *refusal*. The chance of correctly assigning a semantic relation increases if the noun phrase is *parental refusal* and the expression *the parents refused* had already been analyzed.

DIPETT performs some lemmatization. It actually reduces words to their uninflected form. This is the information we use, instead of the results of more general lemmatization.

- **Part of speech**. For complex structures, this will be the part of speech of their head. We use the same parts of speech as the ones used in the syntactic analyzer[3].
- **Syntactic role**. The syntactic role is one of the elements of List 2.

List 2 – Possible syntactic roles

`[subj,`	subject
`complement,`	complement
`attrs,`	attributes
`adverbial,`	adverbial

`np_postmodifiers,`
`pre_modif,`
`post_modif,`
	modifiers for the noun phrase

`s_qualifier,`	sentence qualifier

`rel_clause,`
`single_main_clause,`
`head_main_clause,`
`next_main_clause,`
	type of clause

`initial, final, medial,`	type of subordinate clause

`ing_clause,`
`genitive_ing_clause,`
`to_infinitive_clause]`
	type of relative clause

The elements of this list are also non-terminals from TANKA's grammar. They describe the role played by the syntactic units in the given sentence. The system will assign, as a role to a structure, the first element of List 2 encountered on the path from the structure to the top-most level of the parse tree.

- **Indicator** for the position of a syntactic unit in the sentence (the preposition for a noun which is the head of a prepositional phrase, the subordinator for the head of a subordinate clause, etc.).

[3] For a complete list of parts of speech recognized by the syntactic analyzer refer to Appendix C

The syntactic role and the indicator are grouped in a structure with the following format: *s_role(SyntacticRole,Indicator).*

- **Structure-dependent information.** If the entity extracted is a simple structure (none of the structures we are looking for are embedded in it), extract particular information (for verbs - tense, number, modals, etc; for nouns - number, determiners, etc).

  For an `entity`[4] node, for example, the information extracted has the following format:

```
[PartOfSpeech, StructureType, WordType,
   s_role(SyntacticRole,Indicator),
   Intensifier, Predeterminer, Determiner, Postdeterminer,
   Number]
```

`PartOfSpeech` will be the part of speech of the head of the entity structure; `StructureType` will be `entity`; `WordType` may be `countnoun, propernoun, pronoun`, etc.

If HAIKU does not find any of these elements in the parse tree, it will assign them the value `nil`, as shown in the examples from sentence 1 in the input text:

264 *Clouds tell the story.*

```
cloud [n,entity,countnoun,s_role(subj,nil),nil,nil,nil,nil,pl]
story [n,entity,countnoun,s_role(complement,nil),nil,nil,the,nil,
       sg3]
```

For a `predicate`[5] node, the information extracted has the format:

```
[PartOfSpeech, StructureType, VerbType,
   s_role(SyntacticRole,Indicator),
   s_str(SubcategorizationStructure),
   Tense, Participle, Polarity, Transitivity, Voice]
```

[4] The non-terminal `entity` has a different meaning than the word *entity* that has been used throughout this dissertation. `entity` is a non-terminal that covers a syntactic structure whose head is an *entity* – a person, thing, phenomenon, etc. – which is expressed through a noun or pronoun.

[5] The non-terminal `predicate` covers the syntactic structure whose head is a verb.

`s_str(SubcategorizationStructure)` contains information about the subcategorization structure of the main verb in the `predicate` structure (for example: subject-verb-object ↔svo).

```
tell [  v, predicate, regular, s_role(single_main_clause,nil),
        s_str(svo), tense([present_simple]), partic(nil),
        neg(yes), trans(tr_intr), voice(active)]
```

For other structures, the information will have the format:

```
[PartOfSpeech, StructureType, HeadWordType,
   s_role(SyntacticRole,Indicator)]
```

`HeadWordType` may or may not appear. For pronouns, nouns and adjectives it is instantiated, for adverbs it is not. For example, here are the data for the adverb *up* and the adjective *great* from sentence 5 in the input text:

265 *Far across the pond, great white clouds are rolling up.*

```
up    [adv, adv_cl, s_role(adverbial,nil)]
great [adj, adj, pos, s_role(attrs,nil)]
```

Step 2 **Assign an ID to each syntactic unit**. For open-class words, the ID is based on the word form, the ID of the sentence in which it appears, and the number of times the word was encountered until this point in the text. Complex structures will take the ID of their head. Each structure will be processed only after all structures embedded in it have been processed. An example is presented in Table 5.8.

Step 3 **Add the information extracted for each simple structure to a database, indexed on word forms**.

Each entry in this database has the following format:

```
root(Lemma,LemmaID,Information)
```

`Lemma` is the uninflected form of a word in a sentence. Lemmatization or could be used for further compaction of the representation, by bringing together words from the same family, like *garden* and *gardener* for example.

StructureID	Structure
n([0,1])	n(cloud,countnoun)
entity([0,2])	entity(head_noun(noun(n([0,1]))), number(pl))
n([1,1])	n(story,countnoun)
entity([1,2])	entity(determinatives(deter(the)), head_noun(noun(n([1,1]))), number(sg3))
predicate([2,1])	predicate(regular(verb(tell, tense([present_simple]), neg(yes), trans(tr_intr))), voice(active), complement(svo(entity([1,2]))))
statement([2,1])	statement(subj(entity([0,2])), predicate([2,1]))
simple_sentence([2,1])	simple_sentence(structure(single_main_clause(declarative(statement([2,1])))))

Table 5.8: Syntactic units extracted from the sentence: *Clouds tell the story.*

`Information` is a list whose elements have the following structure:

```
 [SentenceNumber, Counter,
   Syntactic information for this appearance of the word]
```

`Counter` is a number attached to each appearance of the `Lemma` in the text. It is used to discriminate between several appearances of the same word in a sentence.

For example, the noun *story* appears in sentence 1 and sentence 12 in the input text:

266 *Clouds tell the story.*

267 *To them, the clouds tell a story.*

The representation for the lemma *story*, after processing the input text, will be the following:

root(story,1,

```
[[12,3,n,n,countnoun,s_role(complement,nil)],
 [12,4,n,entity,countnoun,s_role(complement,nil),nil,nil,a,nil,sg3],
 [1,1,n,n,countnoun,s_role(complement,nil)],
 [1,2,n,entity,countnoun,s_role(complement,nil),nil,nil,the,nil,sg3]])
```

Step 4 **Build a simplified tree** by replacing all syntactic structures extracted with their IDs, thus resulting in a simplified parse tree. For example, the parse tree for the sentence *Clouds tell the story.* is the following:

```
parse_tree__decla_or_imper(
  simple_sentence(
   structure(
    single_main_clause(
     declarative(
      statement(
       subj(
        entity(
           head_noun(noun(n(cloud,countnoun))),
           number(pl),
           words([clouds]))),
        predicate(
               regular(
                verb(tell,tense([present_simple]),neg(yes),trans(tr_intr))),
               voice(active),
               complement(
                svo(
                 entity(
                    determinatives(deter(the)),
                    head_noun(noun(n(story,countnoun))),
                    number(sg3),
                    words([the,story]))))),
               words([tell,the,story]))))),
      words([clouds,tell,the,story])),
     end_of_input(period))
```

After processing all syntactic units, and replacing them in the initial parse tree, the following reduced parse tree will be obtained:

```
parse_tree__decla_or_imper(
      simple_sentence([2,1]),
      end_of_input(period)).
```

The structure of `simple_sentence([2,1])` is available in the list of syntactic units extracted from the original parse tree.

Step 5 **Traverse the simplified tree and each syntactic structure to extract pairs** that are connected by a syntactic relation (modifier, argument, sub-clause). In terms of the algorithm, the condition of two structures connected by a syntactic relation becomes testing whether one is embedded in the other, or whether they are both on the same level, and are linked by a connective.

The pairs extracted from the structures in Table 5.8 and the simplified tree shown above are presented in Table 5.9.
These pairs are the input to the next stage of semantic analysis – semantic relation assignment. The IDs, through the database of word forms, make the link to syntactic information associated with each structure in a pair.

5.3.1.3 Analyzing Pairs of Syntactic Units

The next stage of processing is semantic analysis of each pair extracted from the parse tree. For the semantic analysis step, the system uses the following resources:

- pairs of syntactic units extracted from the parse tree;

Pair	StructureID	Structure
∅	n([0,1])	n(cloud,countnoun)
∅	entity([0,2])	entity(head_noun(noun(n([0,1]))), number(pl))
∅	n([1,1])	n(story,countnoun)
∅	entity([1,2])	entity(determinatives(deter(the)), head_noun(noun(n([1,1]))), number(sg3))

{[2,1],[1,2]}
(tell,story)
	predicate([2,1])	predicate(
		 regular(
		  verb(
		   tell,
		   tense([present_simple]),
		   neg(yes),
		   trans(tr_intr))),
		 voice(active),
		 complement(svo(entity([1,2]))))

{[2,1],[0,2]} (tell,cloud)	statement([2,1])	statement(subj(entity([0,2])), predicate([2,1]))
∅	simple_sentence([2,1])	simple_sentence(structure(single_main_clause(declarative(statement([2,1])))))

Table 5.9: Pairs extracted from the sentence: *Clouds tell the story.*

- syntactic information about each element in a pair;
- a dictionary of markers (subordinators, coordinators, prepositions, etc.);
- heuristic for relation assignment.

The unit pairs are represented as pairs of unique identifiers that act like pointers. Through them, the system can get access to all information extracted from the parse tree: the head word of the structure, syntactic role, indicators and other information about the structure and the head word (tense, polarity, determiners, etc). Through the head words, the system can access every occurrence of that particular word, in any inflected form in which it appeared in the text. And further, it can access any syntactic structure which had that word as a head word.

Even though we work with pairs of identifiers, in the discussion that follows, we will refer to pairs of words. They are actually, as mentioned above, pairs of pointers to syntactic structures, but referring to them as such would make the discussion unnecessarily complex.

The semantic analysis algorithm is described in Figure 5.3.

```
for each sentence in the input
 {
  for each pair of syntactic units in the sentence
   {
     apply heuristic to find possible semantic relations to assign;
     query the user about the appropriateness of the relations found;
     collect statistics about the system's performance;
     add information about the tagged pair to the knowledge base;
    }
 add information about the network of relations surrounding
                 each noun and verb in the sentence
 }
```

Figure 5.3: Semantic analysis algorithm

5.3.1.3.1 The Heuristic The heuristic used is based on the ones described in (Barker, 1998) and (Delisle et al., 1993). It uses the syntactic indicators available, the syntactic roles of the entities involved, previous evidence of the semantic relations in which each of the entities can be involved and a unified dictionary of syntactic markers. The unified dictionary of markers combines the three dictionaries presented in (Barker, 1998).

The heuristic works in 4 steps. These steps were ordered according to the expected accuracy of the prediction. We justify the accuracy of each step in a discussion of this process which will follow shortly. The system starts with the procedure for step 1, and stops after the first step in which it finds at least one relation to suggest to the user. The heuristic is presented in Figure 5.4. Each step is discussed and explained through examples.

Step 1 - Using relations This step may succeed if some pairs have already been analyzed.

```
for each pair P do
 step1. Find all relations assigned to the same pair of words.
        If a non-empty list if produced, exit the heuristic, and
          start a dialog with the user.
 step2. Find all patterns of semantic relations and syntactic markers
          associated with words in the same part of speech as the
          head word in the pair. Build a similar pattern for the head
          word in the pair under analysis. Match this pattern with the
          ones collected.
        If at least one matching pattern is found, extract the semantic
          relation corresponding to the pair under analysis, exit the
          heuristic and start a dialog with the user.
 step3. The syntactic indicator associated with each word in the pair
          is considered a syntactic marker. It can be a preposition,
          subordinator, coordinator.
        Find all semantic relations associated with the syntactic marker
          of the modifier in the pair, and match the requirements of the
          marker with the information associated with each element of
          the pair.
        If at least one semantic relation was found, exit the heuristic
          and start a dialog with the user.
 step4. Instantiate the list of possible semantic relations to the empty
          list, and start a dialog with the user.
```

Figure 5.4: Heuristic for semantic analysis

Tagged pairs containing the same words as in the pair under analysis are considered the most relevant. We do not expect that the same two words, described by the same attributes, would appear as instances of different semantic relations. Therefore, if a previously tagged matching pair is found, the semantic relation associated with the stored example will be presented to the user as the system's choice.

There are cases when the same two words are connected by different semantic relations, but their attributes (for example, syntactic role) are different. Example:

268 *When* **you look** *at a cloud in the sky, you are looking at millions and millions of these tiny droplets.* (sent. 79)

269 **Look** *at the sky above* **you**. (sent. 2)

Different relations can be assigned to the pair *(look,you)*. In sentence 79 the pair is assigned the Agent relation, while in sentence 2 it is assigned the Direction relation. The word *you* has different syntactic attributes, and they indicate different syntactic cases. In sentence 79 it is the subject of the predicate *look* (and has the nominative case), in the clause *you look at a cloud in the sky*; in sentence 2 *you* is the prepositional complement of the prepositional phrase *above you* (and has the accusative case), which in turn is the complement of the predicate *look*.

On the other hand, if we constrain this step to consider only pairs of structures with the same attributes, the system will not be able to generalize to pairs encountered on different syntactic levels. For example, the system should assign the same AGENT relation to the pairs[6]:

(protest$_1$,student$_1$) extracted from the sentence:

270 *The students protested against tuition fee increase.*

and

(protest$_2$,student$_2$) extracted from the noun-phrase:

271 *student protest against tuition fee increase*

protest$_1$ plays the syntactic role of predicate, and *student$_1$* the subject role. The syntactic role of *protest$_2$* will depend on the sentence in which it appears, but will definitely not be predicate (because it is the head of a noun phrase), and the syntactic role of *student$_2$* is pre-modifier.

In the implementation of this heuristic step we have relaxed the constraint that the attributes that described the elements of a pair must match those of a previously tagged pair. Therefore, this heuristic will sometimes give inaccurate predictions, for example in the case of the pair *(look,you)* presented above. When we reach sentence 79, sentence 2 has already been processed, and the pair *(look,you)* was assigned a DIRECTION relation. The system will find it and propose DIRECTION for the same pair *(look,you)* in sentence 79.

Step 2 - Using patterns The system will reach this step only if in the previous step the system could not find any semantic relations to propose.

This part of the processing is based on the case pattern matching algorithm described in (Delisle et al., 1993). We extended it to treat not only verbs, but also nouns.

[6] The indices serve to distinguish the elements of the pairs

In this step the system focuses on the head of the pair. The head can be either a verb, or a noun. The system will build a network centered on this head. Each edge in the network will connect the head with an element with which it appears in a pair, as a head or modifier. The information on the edge will contain the syntactic role of the units connected, and the semantic relation assigned to the pair (if the pair was already processed).

Consider as an example the sentence:

272 *The graduate students protested against tuition fee increase.*

and the pair *(student,graduate)*.

The system will build the following network, centered on the word *student*:

```
[[student, graduate, s_role(subj,nil),               s_role(pre_modif,nil), _],
 [protest,  student, s_role(single_main_clause,nil), s_role(subj,nil)     , _]]
```

The underscore (_) stands for uninstantiated semantic relations.

For the sentence:

273 *Weathermen watch the clouds day and night.* (sent 11)

and the pair *(watch,weatherman)*, the system will build the following network centered on the word *watch*:

```
[[watch,    weatherman, s_role(single_main_clause,nil), s_role(subj,nil),       _],
 [watch,         cloud, s_role(single_main_clause,nil), s_role(complement,nil), _],
 [watch, day_and_night, s_role(single_main_clause,nil), s_role(complement,nil), _]]
```

Patterns associated with a verb have one additional attribute – the subcategorization structure of the verb: subject-verb-object, subject-verb-object-indirect object, etc. This information is extracted from the parse tree associated with each sentence. In this case, *watch* has the subcategorization structure *svo* – subject-verb-object.

The system will extract, from previously stored patterns, those centered around a word with the same part of speech as the head word in the current pair. For verbs, the system will choose the patterns associated with a verb that has the same subcategorization structure (*svo, svoc*, etc.).

```
for each edge in PC
  find the best matching edge in P;
  eliminate the edge in P ;
 if no matching edges were found, discard P;
```

Figure 5.5: Matching patterns

Each pattern P among the patterns extracted will be matched with the pattern for the current sentence, PC, following the algorithm presented in Figure 5.5.

For example, if the sentence:

274 *Air pilots know that clouds can bring rain, hail, sleet and snow.* (sent 10)

has already been processed, the system finds the following stored pattern:

```
syn_sem_pattern(v,svo,know,
   [[s_role(single_main_clause,nil), s_role(subj,nil),       AGENT],
    [s_role(single_main_clause,nil), s_role(complement,nil), OBJECT]])
```

This pattern has the subcategorization structure *svo*, which makes it a good match for the pattern centered on *watch*.

From the patterns that remain after this filtering, the system will extract the edges that match the pair under analysis, and the relations on those edges. If the list of relations obtained is not empty, it will be presented to the user in the user-dialog stage.

For the examples presented above, we observe that the patterns for the verbs *watch* and *know* match. Since the pair under analysis is *(watch, weathermen)* with the syntactic pattern *[s_role(single_main_clause,nil), s_role(subj,nil), _]*, it will match the edge of the *know* pattern which is tagged with the relation Agent. This relation is put on a list with relations extracted from other matching patterns. This list will be presented to the user in the dialog stage.

If no possible relations have been found, the system goes to Step 3.

Step 3 - Using markers In this step the system will try to use the dictionary of syntactic markers. Each entry in this dictionary contains the following information:

- **Syntactic marker** – can be a preposition, subordinator or coordinator.

- **Relation** – for each relation that the syntactic marker can indicate, there is a separate entry in the dictionary.
- **Polarity** – for some relations (for example Detraction, Prevention), the polarity of the connective, combined with the polarity of the verbs in the clauses it connects, gives a strong indication of the semantic relation between the clauses (Barker, 1998).

At this step the system focuses on the modifier in the pair. If the syntactic indicator associated with the modifier is found in the dictionary of markers, all possible relations signaled by that indicator are collected and proposed to the user in the dialog step.

For example, the processing of the pair *(look,you)* from the sentence:

275 *Look at the sky above you.* (sent. 2)

The modifier *you* is associated with the following information: *s_role(complement,above)*. The syntactic indicator is *above*, and the system finds the following relations in the dictionary that can be signaled by this preposition: LocationAt, Direction. The list presented to the user will be [LocationAt,Direction].

Step 4 - No results The system will perform this step if none of the previous steps has yielded any results. In this step the system does not attempt to find a relation. It will just instantiate the list of possible relations to the empty list, and proceed to the dialog with the user.

This step may seem superfluous. During the actual experiment, this step will set a certain value to one of the parameters that we monitor. It will signal that the system has not found any options to present to the user.

The heuristic presented implements a type of memory-based learning (MBL) – examples of correct classification into semantic relations are stored as they become available, and they are all used to classify a new instance, based on a nearest neighbour algorithm. The distance metric used is given by the goodness of fit (matching) between the pair under analysis and the stored examples.

Using an MBL process ensures that all instances collected are used in the new classification, nothing is omitted. This type of learning has an advantage over other learning algorithms in natural language processing, because in classifying a new instance, it always takes into consideration all previously seen examples.

5.3.1.3.2 Interaction with the User After obtaining a list of possible relations to be assigned to the current pair, the system starts a dialog with the user.

First of all, the user is asked whether the relation presented is acceptable. The user can reject the pair, accept it with a changed order, or accept it as it is.

If the pair has been accepted, the user is presented with the options found by the system. The user may accept a unique relation presented, choose one from a list of several possibilities, or enter her choice.

At the beginning and end of this dialog step, the system notes the time, in order to find the duration of the interaction. The time is expressed in seconds.

The next step is to query the user about the difficulty of the relation assignment to the pair that has just been analyzed.

Figure 5.6 shows a sample interaction with the system. The results have been formatted to fit the page size.

5.3.1.3.3 Collecting Statistics The following values are collected during the system's dialog with the user:

- **order**: the pair presented had the proper order. This value is provided by the user.
- **onus**: the onus of the interaction on the user. This value is provided by the user.
- **time**: the duration of the interaction. The time is computed by the system.
- **number of choices**: the number of possible relations proposed by the system. The number of choices is computed by the system.
- **user action**: the action of the user in the dialog on semantic relations. The possible values are: *accept, choose, supply, discard*. It is automatically computed from the interaction answers and the list of possible relations proposed by the system.
- **heuristic type**: the step of the heuristic process which provided the list of possible relations presented to the user. Possible values: *relation, pattern, marker, no relation*, each value indicates one step of the heuristic procedure. This value is instantiated during the heuristic steps.

These values are used to build statistics of the experiment, to evaluate the system's performance.

```
Analysing sentence number 2     [look,at,the,sky,above,you,.]

Do you accept the following pair and the order: (head,modifier) look you
[reject,,y,n]

look you
[look,at,the,sky,above,you,.]
look    [v,predicate,regular,s_role(single_main_clause,nil),s_str(svo),tense([imperative]),
         partic(nil),neg(yes),trans(tr_intr),voice(active)]
you     [pron,entity,pers_pron,s_role(complement,above),nil,nil,nil,nil,pl]

Possible relations: [lat,dir]
Many possibilities. Enter your choice:
[caus,effe,purp,entl,enab,detr,prev,cooc,freq,prec,tat,tfr,tthr,tto,dir,lfrm,lto,lthr,lat,
 ltd,ornt,conj,disj,acmp,agt,ben,excl,st,prop,poss,posr,inst,obj,obj_prop,recp,part,whole,
 prod,cont,top,cntr,manr,matr,meas,ord,equa,type]      dir

Input estimated onus for this relation assignment:
        0 - obvious relation choice
        1 - reflection required
        2 - serious thought required, but appropriate relation found
        3 - no satisfactory relation exists

[0,1,2,3]       0

Do you accept the following pair and the order: (head,modifier) look sky
[reject,,y,n]

look sky
[look,at,the,sky,above,you,.]
look    [v,predicate,regular,s_role(single_main_clause,nil),s_str(svo),tense([imperative]),
         partic(nil),neg(yes),trans(tr_intr),voice(active)]
sky     [n,entity,countnoun,s_role(complement,at),nil,nil,the,nil,sg3]

Possible relations: [lat,tat,caus,dir,manr,cont,meas,lto,stat,top]
Many possibilities. Enter your choice:
[caus,effe,purp,entl,enab,detr,prev,cooc,freq,prec,tat,tfr,tthr,tto,dir,lfrm,lto,lthr,lat,
 ltd,ornt,conj,disj,acmp,agt,ben,excl,st,prop,poss,posr,inst,obj,obj_prop,recp,part,whole,
 prod,cont,top,cntr,manr,matr,meas,ord,equa,type]      obj

Input estimated onus for this relation assignment:
        0 - obvious relation choice
        1 - reflection required
        2 - serious thought required, but appropriate relation found
        3 - no satisfactory relation exists

[0,1,2,3]       0
```

Figure 5.6: Sample interaction of HAIKU with the user

5.3.1.3.4 Adding Information As the system starts processing the input text, in order to assign semantic relations to pairs of words, it relies mostly on the user, and very little on the heuristic. The reason is that the heuristic has very few indicators to work with, in the absence of previously annotated examples. As the processing advances, the database of semantic relation instances grows, and from the results shown it should be clear that the system relies more and more on the heuristic, thereby reducing the onus on the user.

After processing each pair, the system adds to the database of examples information about the pair and the semantic relation assigned. After processing all the pairs in a sentence, the system adds pattern information for all the nouns and verbs in the sentence. This information will be used in processing new examples.

5.3.2 Running the Experiment

Since the purpose of the experiment is to compare the performance of the system with a previous run, the same methodology and steps of processing must be followed. There are a few differences which do not influence the comparison, and which we will explain. The original experiment is described in detail in (Barker and Delisle, 1996).

Two persons have collaborated in running the experiment in 1996. One ran the system, and the other logged and timed the interaction required. Another reason to have two experimenters was to keep the results more objective, since assignment of semantic relations is quite a subjective task.

The experiment described here also had two subjects. The performance of the system, both from a temporal and interactive point of view, was monitored automatically, by gathering data as the system ran. This data comprised both time and processing statistics. The manual monitoring was not necessary, and the second experimenter helped with keeping the assignment of relations more objective.

As we have already mentioned, there are a few differences.

One of the purposes of this research is to find techniques and resources to improve HAIKU, the semantic analysis module. A better performance of the syntactic analysis module could have a good impact on the performance of HAIKU, since it will deal with better input data.

From the initial version of TANKA, the syntactic analysis module has also undergone some changes (Scarlett, 2000). In this experiment however, for a more objective comparison, we have used the old version of DIPETT. Because of this, it would be superfluous to analyze the performance of the syntactic analysis module. Since in the previous run the output generated by DIPETT was stored, we can bypass the parsing and anaphora resolution phases, and skip directly to semantic analysis.

We use the parse trees generated for the 1996 experiment. Pronoun explicitization was done after generating the parse trees, and before the semantic analysis. A record of the parse trees with anaphora resolution was not kept. We supplied a manually built database of pronouns and their referents. HAIKU will combine the parse trees and the manually built database of pronouns.

The parse trees for some of the sentences were fragmentary. As much information as possible was extracted from them. Some fragments were too small to allow extracting pairs of syntactic units that interact. There were also parse trees that had errors – especially prepositional attachment errors, or errors due to mis-tagged words (e.g. *rain* as a verb was mis-tagged as a noun). The type of errors that DIPETT makes that cause erroneous pairing of syntactic expressions by the new version of HAIKU, are exemplified in sentence 4 from the input text:

276 *Tiny clouds drift across like feathers on parade.* (sent. 4)

HAIKU produces the following pairs:

```
cloud     [n,entity,countnoun,s_role(subj,nil),nil,nil,nil,nil,pl]
tiny      [adj,adj,pos,s_role(attrs,nil)]

drift     [v,predicate,regular,s_role(single_main_clause,nil),s_str(svo),
          tense([present_simple]),partic(nil),neg(yes),trans(tr_intr),
          voice(active)]
feather   [n,entity,countnoun,s_role(complement,like),nil,nil,nil,nil,pl]

drift     [v,predicate,regular,s_role(single_main_clause,nil),s_str(svo),
          tense([present_simple]),partic(nil),neg(yes),trans(tr_intr),
          voice(active)]
parade    [n,entity,countnoun,s_role(complement,on),nil,nil,nil,nil,sg3]

drift     [v,predicate,regular,s_role(single_main_clause,nil),s_str(svo),
          tense([present_simple]),partic(nil),neg(yes),trans(tr_intr),
          voice(active)]
cloud     [n,entity,countnoun,s_role(subj,nil),nil,nil,nil,nil,pl]
```

From these pairs, the pair *(drift,parade)* is wrong, because *parade* should have been paired with *feathers* – *feathers on parade*, and not *drift on parade*, as the syntactic analyzer suggests. This pair will be rejected by the user during the dialog with the system.

Also, one pair is missing – *(drift, across)*. The reason is that *across* is considered the preposition in the fragment *across like feathers*, which is analyzed as a prepositional phrase. *Across* should have appeared as simple adverbial clause attached to the main verb *drift*, separate from the complement *like feathers*.

The new version of HAIKU allows the user to reject an erroneous pair, but it does not give the option of entering a pair that it did not find. The reason for this is the following: if the system could not pair two elements, it means that from the information in the given parse tree, they can not be paired. Even if the user is allowed to input the pair himself, the syntactic characterization of the elements in the pair is still based on the parse tree, and is erroneous. The user should then correct the parse tree, which is a tedious task. With more improvements of the parser, the need for the user to input the pair of entities will disappear.

5.4 Analysis of Results

5.4.1 Learning Analysis

The new version of HAIKU generates pairs of syntactic units that will be tagged with semantic relations. The pairs are treated in the same way, regardless of the syntactic level at which they were found. This causes another difference from the previous HAIKU. At every step only one pair of concepts is tagged with a semantic relation. When processing was separate according to the syntactic level, at the intra-clause analysis step the user was presented with the whole argument structure of the main verb. The system's suggestions covered all verb-argument pairs at once.

It is not clear from the present analysis how individual pair analysis influences the results shown. On the one hand, when analyzing verb-argument pairs one by one, even if not all relations are correctly suggested, the system can get some points for individual good proposals. On the other hand, the system is penalized only with one mistake when it gives the wrong suggestion for the entire network of relations around the main verb. In individual analysis, the system will be penalized with a wrong answer for each relation, thus lowering its overall score.

For example, in processing the first sentence of the test text:

277 *Clouds tell the story.* (sent. 1)

the original version of HAIKU will propose patterns that cover the entire verb-argument network. It should propose patterns of the following kind:

agt-obj

number of analyzed examples	1465				
level statistics	CL	IC	NP		
	64	974	427		
user actions	accept	choose	supply		
	451	402	612		
average number of suggestions	2.81				
user onus	0	1	2	3	average
	82.05%	12.07%	5.25%	0%	0.18

Table 5.10: Summary of semantic analysis

expr-obj
...

Since it is the first sentence in the text, the system has no suggestions, so the user will input the pattern **agt-obj**, and the system will be penalized with one error.

In the current version of HAIKU, the user is presented with verb-argument pairs one by one. In this case *(tell,cloud)* and *(tell,story)*. Again, since it is the first sentence to be processed, and there are no syntactic markers to be used, the system has no suggestions. There will be two interactions with the user with no suggestions from the system, so it will be penalized twice.

Table 5.10 presents some statistics of the pairs extracted, and the results of the processing. Out of 2020 pairs extracted from the parse trees, 555 were discarded by the user during interaction. We will show the statistics for the accepted pairs.

Since the number of suggestions when the user accepted a unique relation proposed by the user is obviously 1, the average number of suggestions was computed only for the cases when the user chose one of the relations suggested by the system, or provided her own after discarding the possibilities presented. The reason is that including the accepted answers in the average would artificially lower this number.

In a manual comparison of the output produced by the old and the new version of HAIKU, we have noted that the new version of HAIKU processes some of the sentences that the old version skipped. In the old HAIKU only clauses around a non-stative verb were considered. The sentence:

278 *The cloud is all around you.* (sent. 60)

was not processed. In the new version of HAIKU we choose to analyze this type of sentences as well.

Clause fragments will also be processed. For example, the following sentences and the pairs extracted:

279 *Always floating.* (sent. 25)

```
float   [n,entity,ing_noun,s_role(emb,nil),nil,nil,nil,nil,sg3]
always  [adj,adj,pos,s_role(attrs,nil)]
```

280 *All kinds of clouds.* (sent. 49)

```
kind   [n,entity,countnoun,s_role(emb,nil),nil,nil,all,nil,pl]
cloud  [n,entity,countnoun,s_role(np_postmodifiers,of),nil,nil,nil,nil,pl]
```

The overall results are shown in Figure 5.7. The plot shows the cumulative number of examples accepted, chosen, supplied and either accepted or chosen by the user during the processing of the input text. Figure 5.8 shows a zoom in on the first 60 pairs processed, in order to analyze better the behaviour of the system towards the beginning of the experiment.

The system starts to propose good options to the user by the 8th pair analyzed. The pair is *(drift, feather)*, in the sentence:

281 *Tiny clouds drift across like feathers on parade.* (sent. 4)

The step of the heuristic that gave results was step 3 – using syntactic markers, in this case *like*. The system proposed more than one relation, and the user action was *choose*.

The first *accept* action occurred for the pair *(cloud,white)*, in the sentence:

282 *Far across the pond, great white clouds are rolling up.* (sent. 5)

The system used a previously stored pattern to find a possible relation, which was also the correct one.

We observe that the system's performance is improving as the text is processed. The curve that shows the combined results of the *accept* and *choose* user actions surpasses the curve that shows the cumulative number of *supply* actions. This means that as the system accumulates processed examples, it makes better and better suggestions. Most of the time the user will be in a situation of accepting or choosing one of the system's suggestions, instead of providing an answer.

Figure 5.9 shows the usage of the heuristic type over the course of the experiment. The type of heuristic used is extracted only for the examples for which the system provides the correct answer (the action of the user is either *accept* or *choose*).

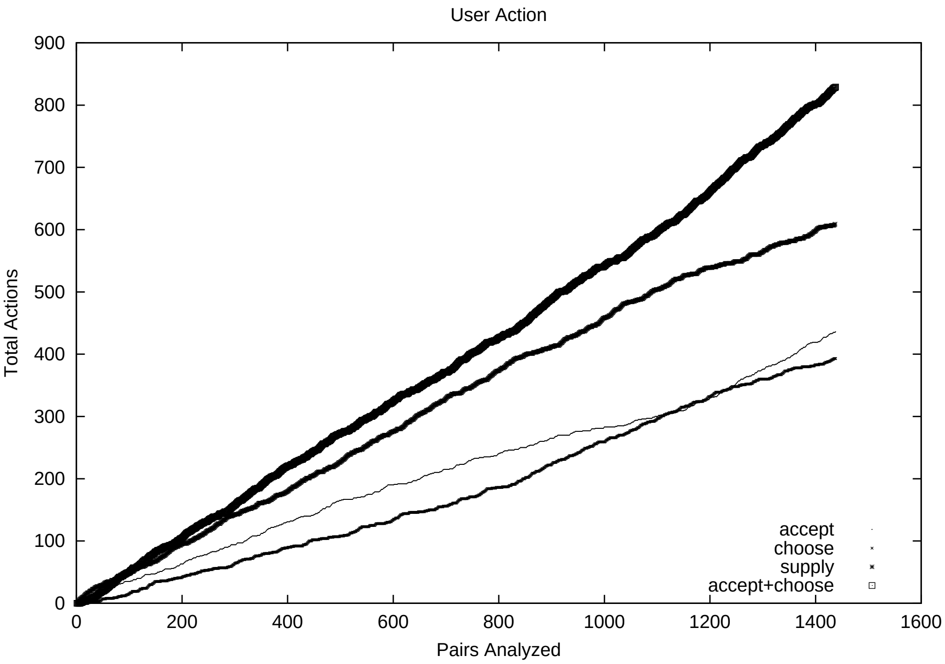


Figure 5.7: User action results

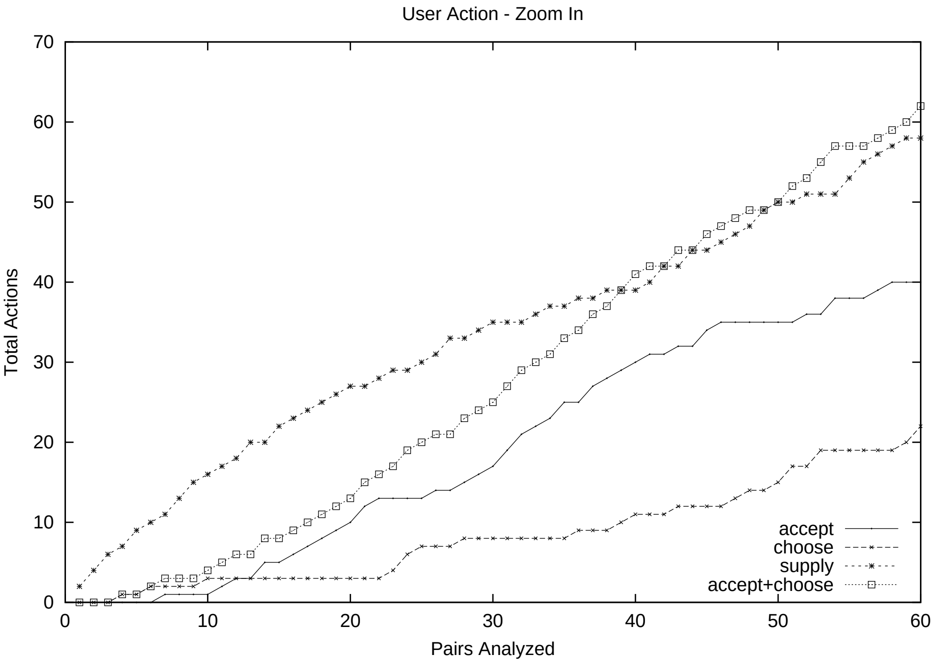


Figure 5.8: User action results - zoom in

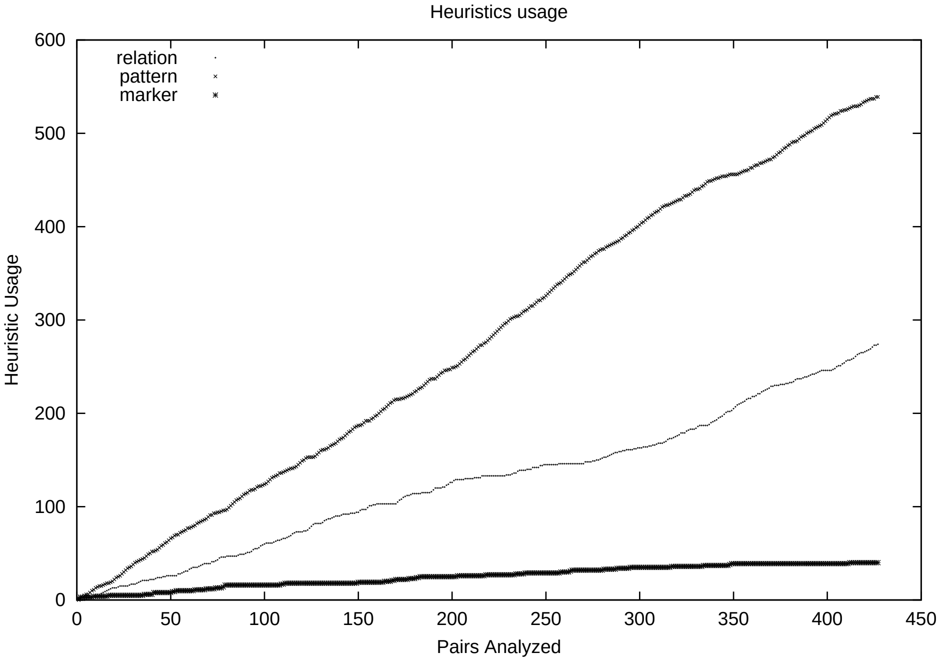


Figure 5.9: Heuristic statistics

The usage of patterns in finding possible relation assignments surpasses the other two heuristic steps. A possible reason may be the fact that there are more verb-argument pairs (intra-clause level pairs), than noun-modifier or clause-clause pairs (974 versus 427 or 64, as shown in Table 5.10). This is indeed the case. As seen in Figure 5.9, for assignment of relations for verb-argument pairs, the pattern heuristic step will provide the correct answer in most of the situations.

We will show plots built separately for three syntactic levels – clause, intra-clause and noun phrase. The system does not treat pairs differently according to the syntactic level they come from. Level information was kept only for comparison purposes. The original system analyzed each level separately, and we want to be able to produce parallel statistics for better comparison.

Compared with the performance of the original HAIKU, the new system starts learning earlier. When processing case relations, the original system processed half the input examples before the combined results of the *accept* and *choose* actions surpassed the *supply* one.

For comparison, Figures 5.10 and 5.11 show the behaviour of the original and the new system, only for relations at the intra-clause level (cases).

It is interesting to notice that the performance of the system only for pairs extracted

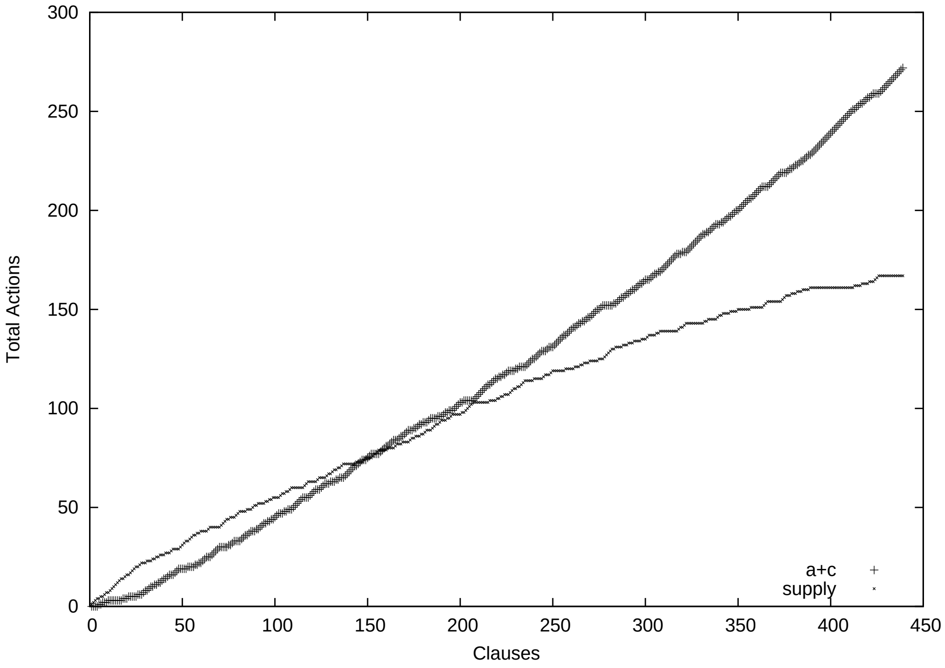


Figure 5.10: Original HAIKU: User action over time for case assignment

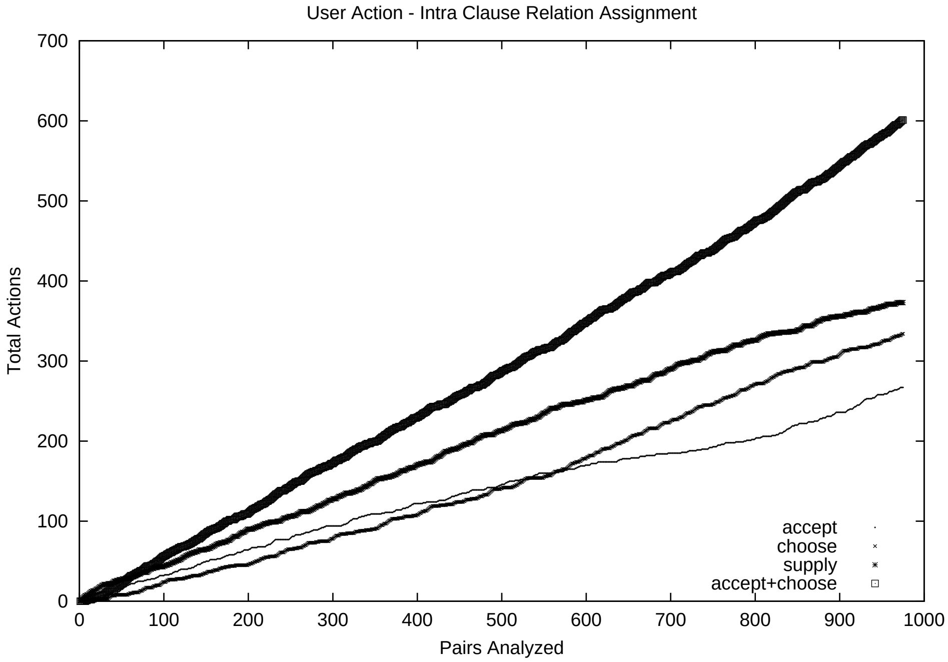


Figure 5.11: New HAIKU: Statistics for the intra-clause level

at the intra-clause level closely mirrors the overall performance of the system. This can be observed by comparing the statistic on the heuristic steps for the overall performance and for pairs at the intra-clause level.

Figure 5.12 shows the usage of the heuristic steps in assigning cases for the intra clause level. The system relies mostly on finding a matching pattern.

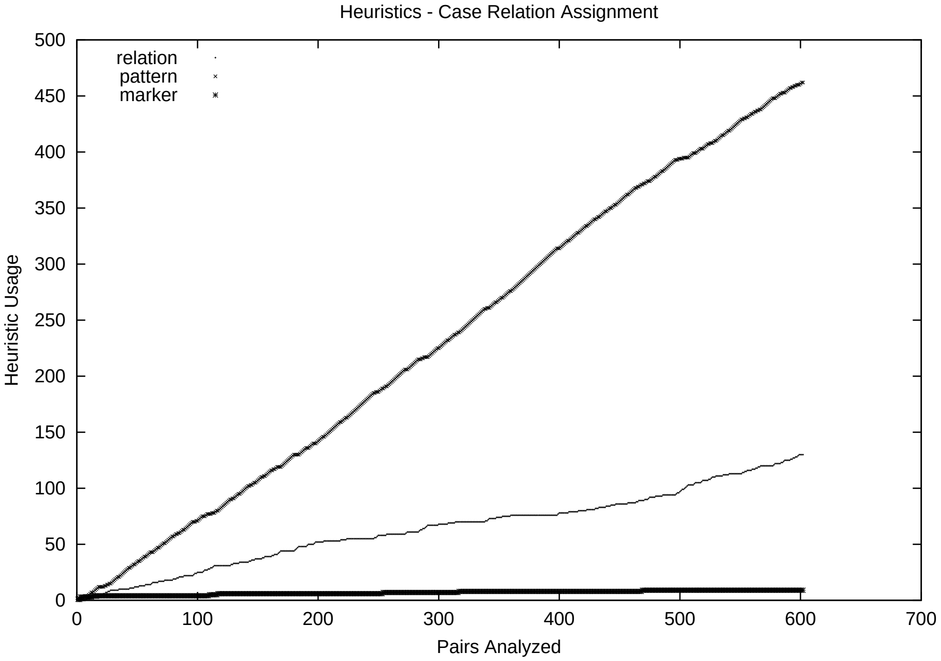


Figure 5.12: Heuristic usage for case assignment

The plots for the user interactions in assigning semantic relations in noun phrases looks a bit different than the one for case analysis.

It is interesting to observe that the plot that shows the cumulative number of examples accepted by the user becomes steeper as the experiment progresses. The text processed is a semi-technical text on meteorological phenomena, written in a very simple style. Based on this, an explanation of the increasing performance of the system as more examples are processed may be the fact that many noun phrases are repeated throughout the text. We also observe the increasing slope of the curve that shows the number of examples for which the user must supply a relation. An explanation of this fact could be that all through the text new concepts, expressed as noun phrases, are introduced. When they are first encountered, the system cannot make a good suggestion. If this is the case, an analysis of the heuristics used to provide good suggestions at this syntactic level should show a predominance of the first step of the heuristic (based on retrieving the relation

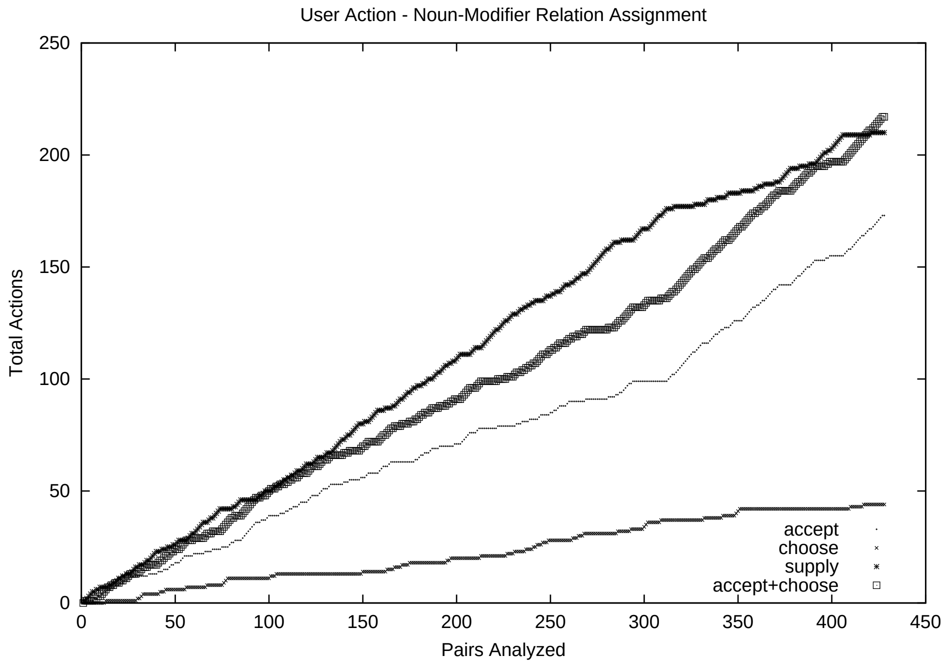


Figure 5.13: New HAIKU: Statistics for the noun-phrase level

associated with pairs consisting of the same words as the pair under analysis).

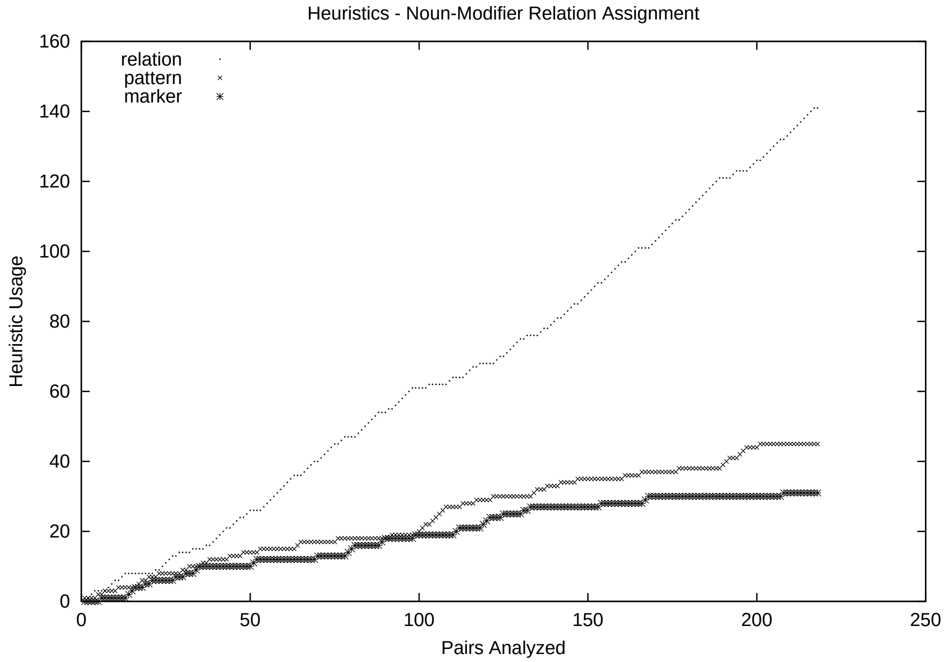


Figure 5.14: Usage of the heuristic steps in assigning NMRs

At the beginning of the experiment, the system makes use of patterns and syntactic markers to boost its performance. The noun phrases in the input text do not have a complex structure. The simpler they are as a structure, the less use the system can make of syntactic markers. The usage of both markers and patterns levels off, and the system will rely mostly on stored pairs and on the user.

A surprising behaviour comes from the clause-level relation assignment. From the analysis of the heuristics used, it becomes evident that using patterns yields some possible relations. Therefore, step 3 of the heuristic is not reached. Step 3 is the one where syntactic markers are used, and CLR assignment is mostly based on them (Barker, 1998). This is the side effect of the ordering of heuristic steps. By this ordering we have tried to force the system to learn. The first two steps rely on the information that the system has accumulated, whereas step 3 relies on a more static type of information – the dictionary of markers which is modified only when a new marker or an old marker indicating a new relation is encountered.

If that is the case, we expect the system to perform worse than the original system, in which CLR assignment relied almost exclusively on syntactic markers. Figure 5.16 shows the statistics for user action for CLR assignment. The system does not perform as well as for the other levels.

Table 5.11 shows the relation between the heuristic and the user's actions. When the user action is *accept*, the suggestion made by the system is mostly based on stored examples that consist of the same word as the current pair. Patterns also provide good answers in about 40% of the situations when the user accepts a unique suggestion by the system.

	relation	pattern	marker	no suggestion
accept	269	182	0	0
choose	5	357	40	0
supply	45	394	22	151

Table 5.11: Heuristic - user action relation

Step 1 of the heuristic (Section 5.3.1.3.1) explains why it is possible to have a *choose* or *supply* user action when the system uses a previously stored pair with the same words as the present pair. We have relaxed the matching between the attributes that describe the elements in two pairs that are compared (syntactic role, indicators, etc.), so that pairs extracted from different syntactic levels, which obviously will have different syntactic roles (at least), could possibly match.

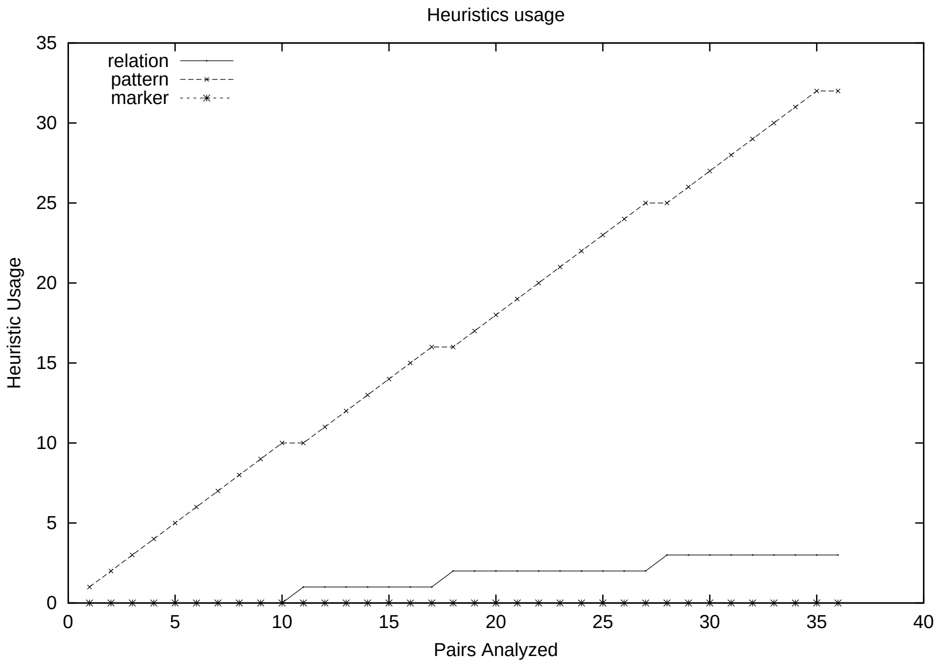


Figure 5.15: Heuristic usage for CLR assignment

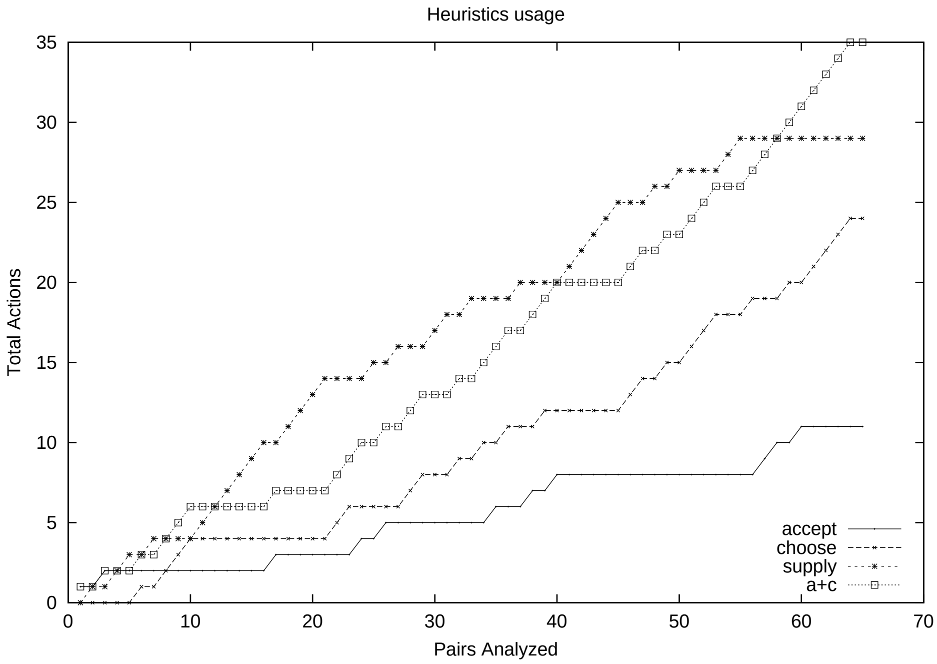


Figure 5.16: User actions for CLR assignment

Step 2 of the heuristic, using patterns, is the most productive. This, as suggested in the discussion on case assignment, is due to the prevalence of verb-argument pairs analyzed by the system. For case assignment, patterns perform best.

5.4.2 Onus on the User

We have computed two measures that show the burden that the new system puts on the user: time of interaction and onus on deciding the semantic relation to assign a pair.

The experiment was performed during 5 sessions of approximately three hours each, due to the schedule of the experimenters. The overall time spent on semantic relation assignment is 6 hours, 42 minutes and 52 seconds. The time was computed by adding the time recorded by the system for each interaction. The timing was started when the system presented its suggestions to the user, and was stopped when the user entered their choice.

The two subjects who have conducted this experiment have recorded an onus value for each of the interactions. The values range from 0 to 3:

0 – obvious relation choice (1202);

1 – reflection required (186);

2 – serious thought required, but appropriate relation found (77);

3 – no satisfactory relation exists (0).

There were no pairs to which we could not assign a relation, there were however 77 which were tagged after serious considerations of the available options.

The average onus on the user was 0.18. The previous TANKA system reports an average of 0.11 for the clause level and 0.14 for the intra-clause level analysis (Barker and Delisle, 1996).

A possible explanation for the difference in user onus in the two experiments may lie in the fact that in the experiment described here, one of the subjects had only a brief experience with the semantic relations before the experiment took place. In the original experiment, both subjects had been working on the project for a long time. Another issue is that in combining the three lists of semantic relations that the system originally used, we have split more general relations on one level into several, as indicated by more specific relations on other levels. It means that the combined list is more fine-grained that the three lists that it combines. This may have added to the difficulty of the task, since more fine distinctions had to be made in order to decide on a particular relation.

5.5 Conclusions

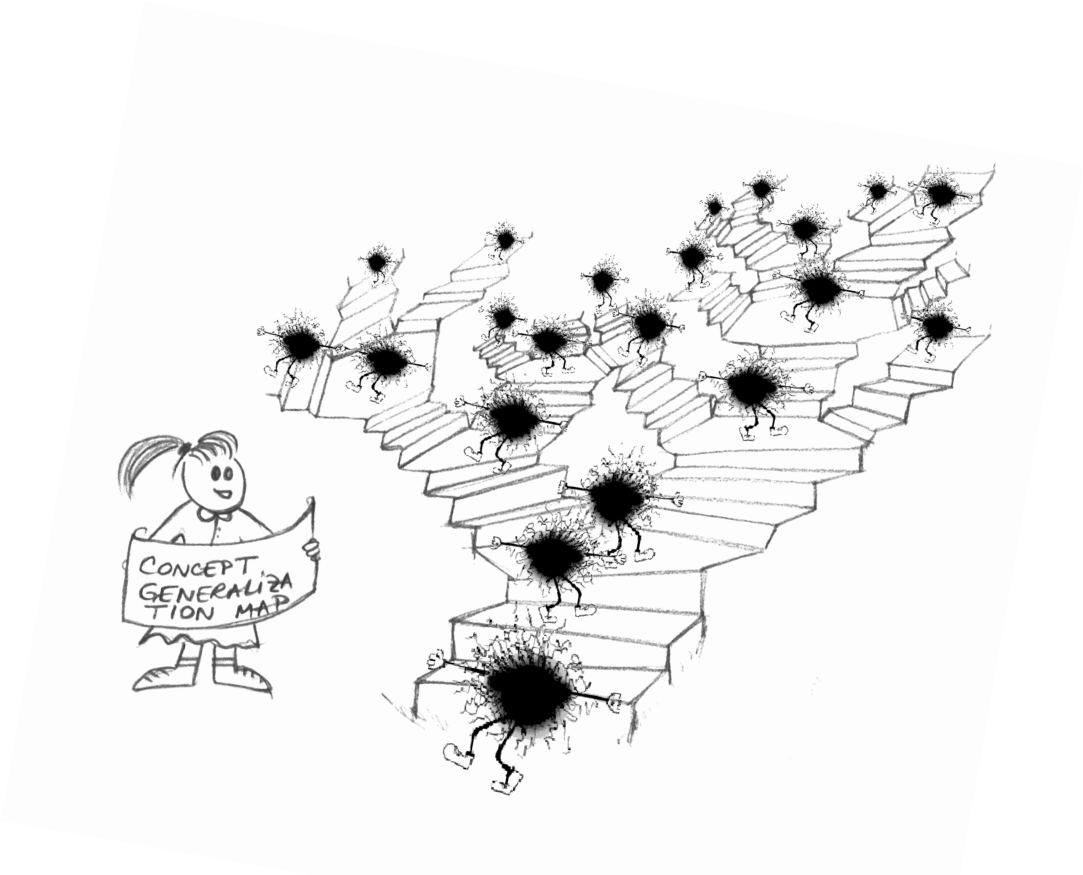
Through the comparative analysis of the results of our experiment conducted using a combined list across three syntactic levels, and the experiment conducted using three separate lists, we observe that our version of the system works better. When it uses a combined list, the system starts making good suggestions by the fourth sentence analyzed. After analyzing approximately 10% of the input (50 sentences out of 513), the number of good suggestions that the system makes surpasses the number of erroneous suggestions. In the original system almost 40% of the input had to be analyzed (200 out of 513 sentences) before the number of good suggestions exceeded the number of erroneous ones.

It is difficult to properly evaluate such a system with only two judges. The task of semantic relations assignment, as most other semantic processing tasks, is highly subjective. We would need to perform the same experiment using several judges and then compare the results, as well as the agreement between these judges. The resources at our disposal did not allow us to perform such a large-scale experiment. We have kept our results somewhat objective by bringing in one experimenter that had no bias towards any kind of semantic analysis. She was presented with our list of semantic relations, with examples and definitions. She performed the task of assigning semantic relations based only on her understanding of the message in the sentence, and the available labels to describe the interaction of concepts. The other experimenter had knowledge of the system and the semantic relations, and was present to perform the dialog with the system, and be available for general consultations on the semantic relations to be assigned.

Further experiments with several judges will be done in the future.

Chapter 6

Exploring Semantic Relations Using Ontologies



6.1 Introduction

We have presented in Chapter 1 the idea that motivates our research: since there is a limited number of semantic relations that link concepts and a virtually unlimited number of pairs of concepts that interact, pairs of concepts connected by the same relation must have something in common. We have investigated until this point the link between different syntactic manifestation of concepts. This has lead us to a unified view of semantic relations across syntactic levels, whose computational benefits were explored in Chapter 5.

In this chapter we will explore, computationally, similarities between concepts using ontologies. If by using lexical resources we find that the concepts connected by the same semantic relations can be clustered because of some semantic characteristics that they share, we would like to incorporate this in our text analysis system.

If while analyzing a pair of concepts we find that these concepts share certain characteristics, this will allow us to find quickly, based on pairs that have already been analyzed and share the same characteristics, the semantic relation that best describes the way these concepts interact. This will help the knowledge acquisition process, allowing the system to rely more on previously analyzed examples, instead of relying on the user.

Throughout the theoretical investigation on semantic relations, we have adopted a coarse view of concepts which distinguished only among occurrences, entities and attributes. If we want to identify semantic relations based on semantic characteristics of the concepts they connect, such a view is not adequate.

We have seen that beside research that is concerned with assigning semantic relations to pairs of syntactic units, research in knowledge representation assigns semantic relations to pairs of concepts. Semantic relations are defined in terms of conceptual primitives that characterize the concepts these relations connect (Sowa, 1984). In order to implement this approach in a computational text analysis system, we require a resource that describes every concept in terms of conceptual primitives. Because of controversies surrounding theories of concepts as collections of primitives, such a description may not even be possible.

We can attempt to find automatically some description of concepts in terms of semantic features they share by using machine-readable ontologies of concepts.

WordNet (Miller et al., 1995) is one of the most extensively used resources in the computational linguistics community. *WordNet* is used in annual word-sense disambiguation competitions (SENSEVAL), question answering and document understanding competitions, it has spawned versions in other languages (EuroWN), and there are conferences that focus exclusively on issues surrounding it. The tasks for which *WordNet* is used may uncover shortcomings, which are addressed and then improved versions of *WordNet* are released to be freely used by the research community.

Roget's Thesaurus (Kirkpatrick, 1987) and *LDOCE* (Procter, 1978) are resources that are gaining popularity with the research community in NLP. *WordNet* has the advantage that it is free, while *Roget's Thesaurus* and *LDOCE* require a commercial license.

The following intuition guides these experiments: since we have many pairs of concepts (A_i, B_i) whose interaction is described by a semantic relation R, then maybe the concepts A_i have all something in common, and the same for B_i. In order to explore such possibilities, we need to represent these concepts in a semantic space, whose dimensions give us their description. Then we can search for aspects that bring concepts together, and that can explain why concepts from various groups interact with each other in the same way.

The lexical resources available can provide the type of semantic space that we look for. The IS-A links in *WordNet*, and *Roget's Thesaurus*'s organization of knowledge, give us ontologies we can use to try to group concepts. We expect to group concepts by generalization.

In order to find groupings in the resources we use, we employ machine learning techniques, which we describe in Section 6.2. We experiment with finding rules that describe the types of concepts that are involved in specific semantic relations. We use the combined list of relations presented in Section 5.2 to assign semantic relations to the pairs that we analyze. We focus in these experiments on $(noun, modifier)$ pairs extracted from base noun-phrases[1]. These experiments are presented in Section 6.3.

6.2 Using Machine Learning Techniques to Gain Insight Into the Nature of Semantic Relations

The list of relations described in Section 5.2 will be used in a knowledge acquisition system, to indicate how pairs of concepts extracted from texts interact. We are curious about the nature of these relation. We would like to find reasons why a set of examples is grouped under the same semantic relation.

In order to use statistical analysis for this problem we require a large corpus annotated with semantic relation information. Since such a resource is not available, we use machine learning tools which can extract regularities from a comparatively smaller amount of data than statistical tools.

The field of machine learning provides techniques and tools for finding patterns in the

[1] Base noun-phrases are noun phrases which consist of only a noun and one modifier.

data; the patterns found can then be used to analyze novel situations (Mitchell, 1997).

We also seek patterns that explain the grouping of our examples under different semantic relations. We represent our data in a semantic space whose dimensions are given by attributes that we consider relevant with respect to semantic relations. Through these patterns we hope to achieve a better understanding of the way semantic relations are characterizable in terms of syntactic indicators and information extracted from lexical resources.

Because we will use the patterns to gain insight into the nature of semantic relations, the patterns obtained should be in an easily understandable form. For this reason we choose to experiment with machine learning tools that perform symbolic learning, like decision tree or rule induction systems, as opposed to the more opaque neural network or support vector machine systems, for example.

We will present in the sections that follow the techniques we choose to experiment with, and the actual implementations of these methods we use.

6.2.1 Decision Trees

Decision tree learning consists in finding descriptions of the classes to be learned by building a tree. In each node there is an attribute, according to whose values the data is split into subsets. These subsets will be further split according to the values of other attributes, until the subsets obtained are pure enough with respect to the classes of the data (according to some parameter), a preset depth has been reached, or the tree overfits the data (the tree built is too specific, and matches too closely the input data; it does not generalize well based on the attribute values that describe the data).

The choice of an attribute for a given node is determined by its *gain ratio*. The gain ratio shows the proportion of information generated by the split that is useful for classification.

We have used the C5.0 implementation of this particular ML technique (RuleQuest, 2000), and we will use the following parameters:

- **Training and testing**. If desired, the input data can be split into a training and a test set. The training set is used to build a decision tree whose performance is measured using the test set. The split preserves the ratio of examples across the classes.

- **Cross-validation**. A parameter can be set to the N number of cross-validation sessions to be performed. The data set will be split into a number N of subsets. At each turn, one of the subsets will serve as a test set, the rest will be used for training. The split preserves the ratio of examples across the classes.

- **Misclassification costs**. It may be the case that one of the classes represented in the data set is more important. In order to give more emphasis to correctly identifying that particular class, C5.0 gives the option of specifying misclassification costs – the system will be penalized more severely for misclassifying examples in the class that is more of interest. Misclassification costs are also used in situations where there is a big difference among the number of examples in each class. This situation is called *class imbalance*. C5.0 is sensitive to this problem.

- **Rule sets**. Instead of obtaining the output as a tree, this parameter gives the option of viewing the results as a set of rules. They are obtained by traversing the decision tree, the path will be the condition, and the leaf gives the class. They can be simplified by eliminating conditions that do not contribute to the discrimination of the nominated class from other classes.

C5.0 offers other parameters as well, which allow it to build or modify the decision trees in different ways (for example, *pruning*), or modify the learning process (*boosting*). The ones described are the only ones we will use.

6.2.2 Rule Induction

The idea behind this technique is to iteratively induce rules for each class's positive instances. Each rule is represented as a disjunction of conjunctions. The method of search is "divide-and-conquer" – if each example is regarded as a point in a multidimensional space (the dimensions of the space are the attributes that describe the examples), rule induction is based on splitting the space into regions, and building a rule that characterizes the examples in that region.

Rule induction systems have the desirable property of producing results that are relatively easy for people to understand (Catlett, 1991). In comparative studies with decision tree systems it was observed that rule learners perform better on many problems (Pagallo and Haussler, 1990), (Quinlan, 1987), (Weiss and Indurkhya, 1993). On the downside, Cohen (1993) shows that they scale poorly with the sample size, especially when the input data is noisy.

RIPPER (Repeated Incremental Pruning to Produce Error Reduction) is a tool that implements the rule induction learning technique (Cohen, 1995). It also solves some of the problems of previous rule induction systems – it reduces the error and increases the scalability of the system to the size of the input data. It is based on a repeated grow-and-simplify approach that generates a set of rules, and each rule obtained is post-processed to minimize the error of the whole set. For each rule a few alternatives are built, which are then tested in the context of the whole set. The one that performs best in this context is kept.

Parameters used:

- **Classes to learn**. As opposed to C5.0 which gives the same importance to each class it tries to learn, RIPPER can focus on N-1 out of N classes it will learn. In the case of a binary classification problem, the system has the option of producing only rules that characterize the positive class.

- **Misclassification costs**. In the presence of imbalance, the system's performance can be changed by assigning a misclassification cost to force it to focus on the class that is most interesting for the given experiment.

Using misclassification costs did not have much influence on the outcome of our experiments. An initial attempt to classify the input data into 30 different classes (corresponding to the semantic relations represented in the data set) gave poor results. We have changed the 30-class learning problem into 30 binary classification problems. The first option mentioned, in which the system will focus only on the positive class, has eliminated the need for misclassification costs. The only effect misclassification costs had was to produce longer rule sets.

6.2.3 Inductive Logic Programming

Inductive logic programming (ILP) is based on representing theories as logic programs. It combines inductive methods with first-order logic representations to produce general first-order theories from the input examples (Muggleton and Raedt, 1994). It can learn relational knowledge that cannot be captured by systems that use attribute-value methods.

FOIL is a particular implementation of ILP (Quinlan, 1990). It employs a top-down learning method. The system starts with a very general rule, which is iteratively specialized until it becomes consistent with a subset of the positive part of the input data.

This subset of positive examples is then removed, and the process of building a rule is performed again until all positive examples have been covered.

The output consists of rules, actually Horn clauses. After generating clauses that cover all positive examples, they are filtered to eliminate possible redundancies. The final set of clauses is reordered so that recursive rules come after the corresponding base-cases.

From a method of learning point of view, FOIL is quite similar to decision trees. The difference comes from the fact that decision trees test attribute values, whereas ILP tests first-order literals.

6.2.4 Memory-Based Learning

Memory-based learning (MBL), also known as lazy, instance-based, case-based, k-nearest neighbour learning, is a technique that does not compute or store explicit abstractions (rules, trees, etc.). Instead, it stores all the instances of the input data. At classification time it predicts the class of previously unseen examples based on similarity measures between these examples and the stored instances (Cover and Hart, 1967).

This type of learning has some limitations: it requires space for storing the whole input data; also, it does not create descriptions for the classes it tries to learn.

There are also advantages to this type of learning. One is the fact that it is quite a fast algorithm. The other, of more importance to the NLP community, is the fact that it is capable of dealing with outliers in the representation space. Daelemans et al. (1999) show that exceptions are very important for NLP, and MBL (as opposed to other learning techniques) is the one that can take them into consideration.

We will not use an off-the-shelf tool in this case. We have implemented a task-specific distance measure, and we apply the 1-nearest neighbour algorithm.

6.3 Testing the Use of Lexical Resources

If a person assigns the same semantic relation to different pairs of concepts, a legitimate question that comes to mind is whether the pairs have something in common. One possibility may be that either the first, the second, or even both elements of each pair are grouped together in a semantic space. Another possibility may be that projecting the pairs onto some semantic space, the relative distance between elements in a pair is consistent over pairs tagged with the same semantic relation.

The semantic space we can use is given by lexical resources. We choose to use *WordNet* 1.6 and *Roget's Thesaurus*, two resources that are available to us. *WordNet*'s and *Roget's*

Thesaurus's IS-A hierarchy define two ontologies onto which we project the word senses in our noun-modifier pairs. We test which of these resources is a better fitting semantic space with regards to semantic relations. Based on the results obtained, we will decide whether either of these resources will be part of the new HAIKU, and TANKA.

The most common methods of assessing word similarity compute a distance (Budanitsky and Hirst, 2001), or find the information content of the most specific subsuming concept in the IS-A hierarchy in a lexical resource (Resnik, 1999). Similarity between noun-modifier pairs is more complex. The distance between the two heads or two modifiers in a pair of base NPs can be zero, yet there may be no similarity between them, with respect to semantic relations. For example: *snow blindness* - EFFECT versus *snow report* - TOPIC (sense 2 of the noun *snow* in *WordNet 1.6* - {*layer*}) or : *pressure cooker* - INSTRUMENT versus *heavy cooker* - PROPERTY (sense 1 of the noun *cooker* in *WordNet 1.6* - {*cooking utensil, cookware*}). There may be a way of combining the semantic distance between the heads with those between the modifiers. We are not looking for such a formula. Instead, we use the same lexical resources that are employed in finding distance metrics, and we extract features that characterize the words in base NPs. There may be several reasons why base NPs are similar. The similarity may be between the components of the base NPs, for example: AGENT – *student protest* and *animal attack* - the modifiers are sentient beings, the head nouns express actions. Otherwise, there may be a relational similarity, for example: TYPE – *oak tree* and *cumulus cloud* - in both NPs the head noun is a hypernym of the modifier. Each relation may have its own signature, as far as such characteristics as described above are concerned. We will therefore let the machine learning tool find the appropriate combination of attributes for the purpose of characterization.

Several attempts have been made to learn the assignment of semantic relations to modifier-noun pairs, without necessarily seeking insight into their nature. The domains, the lists of relations and the methods all vary.

Rosario and Hearst (2001) perform ML using neural networks. They learn semantic relations between a noun and its modifier in a medical domain, to which the list of semantic relations and the lexical resource have been tailored.

Rosario et al. (2002) present a continuation of that research. The authors look manually for rules that classify correctly noun compounds in the medical domain, based on the MeSH lexical hierarchy (Medical Subject Headings). The noun compounds are extracted automatically, and sampled for manual analysis. The hierarchy is traversed in a top down manner to find a level at which the noun compounds displaying different relations are properly separated. Analysis has shown that finding the appropriate level of generalization depends on the relation involved; some are easier to capture in rules than others.

Vanderwende (1994) uses a dictionary built from texts to find clues about possible semantic relations in which the word might be involved (for example, finding *for* in some definition indicates that, in combination with another word, it could display the Purpose relation). In this work words are taken one by one, with no interest in generalization.

For general NPs, Barker and Szpakowicz (1998) use a simplified case of memory-based learning. They store noun-modifier-indicator-relation tuples (the indicator is usually a preposition), and match a new NP with previously stored patterns. No lexical resource is used.

In an experiment that does not involve modifier-noun pairs, Li and Abe (1998) generalize case frames of specific verbs to concepts using *WordNet's* ontology. The experiment aims to find generalizations for the fillers of each syntactic argument of a specific verb, by finding an appropriate cut in the tree structure (defined by the hypernym/hyponym relations in the resource) that covers the examples extracted from a corpus. The best of several possible cuts in the tree is chosen according to the minimum description length principle.

Clark and Weir (2001) present a similar approach in choosing the sense of a noun in *WordNet*. The choice is constrained by the predicate whose argument the noun is, and by the probability of the semantic class to which the noun can belong according to its senses in *WordNet*.

Lauer (1995) maps words in noun compounds onto categories in *Roget's Thesaurus*, in order to find probabilities of occurrence of certain noun compounds and their paraphrases. There is no automatic process in finding the best level of generalization.

All these approaches consider the generalization level of *one* concept. In this process, only words are used. Our approach is different. We look at generalizations of two connected concepts. Different methodologies, data and resources used in the experiments in the approaches we mention. A direct comparison of results is therefore not possible. We believe, however, that our approach is competitive: we analyze in parallel the two concepts involved in a relation, and we investigate general (noun,modifier) pairs, not restricted to a specific domain. This is reflected in the lexical resources we use, *WordNet* and *Roget's Thesaurus*, which are not domain-specific.

There are several features which preliminary analysis has shown to be relevant to recognizing the relation between the concepts: is any of the words the result of nominalization or adjectivalization, is it an -er nominal, is it a noun, adjective or adverb. The aim is to find rules which justify the existence of certain type of interaction between the two elements of the base NP, through the analysis of information extracted from publicly available resources for a general domain, more general semantic relations, and ML methods that present an insightful look into the nature of the data.

The input data is a set of 600 modifier-noun pairs. The modifiers are nouns, adjectives or adverbs. These examples were gathered manually from (Levi, 1978), automatically from (Larrick, 1961), semi-automatically from *SemCor* (the version annotated with *WordNet 1.6* senses). Some examples were constructed and added for relations for which few or no examples were found in these texts. The examples that were not extracted from *SemCor* were manually annotated with *WordNet 1.6* senses. All the pairs were manually annotated with 30 semantic relations from our set of 50. Preparation and processing of the data is described extensively in (Nastase, 2001).

We have performed word sense disambiguation in *Roget's Thesaurus* using *WordNet*, to obtain the equivalent set of pairs annotated with *Roget's Thesaurus* senses (Nastase and Szpakowicz, 2001). Because the two resources do not cover the same subset of the English language, the input data set was reduced to 550 pairs.

The final data set was obtained by representing each modifier-noun pair using the information provided by the ontology underlying *WordNet*. Because *WordNet*'s IS-A hierarchy is a graph, not a tree, one word may have several representations using this information (we will explain this further in Section 6.3.2). Because of this issue, the final number of examples is 767, and 556 after filtering the word senses that do not appear in *Roget's Thesaurus*. We present the distribution of semantic relations in the final data set in Table 6.1.

In order to discover regularities that support the distribution of the data into classes determined by the semantic relations we have manually assigned, we use this data in machine learning experiments with various tools. The choice of a learning tool is driven by the type of results expected – symbolic rules, easy to understand – and by the type of processing of the attributes imposed by our task – generalization in the ontology that underlies *WordNet* or *Roget's Thesaurus*.

We experiment with several learners.

Table 6.1: Distribution of semantic relations in the data set before/after filtering with *Roget's*

Rel	Occur	Rel	Occur	Rel	Occur
cause	19/15	agent	73/35	whole	10/7
effect	37/31	beneficiary	11/9	product	20/16
purpose	44/31	object	45/27	source	21/9
detraction	4/4	object-property	15/13	content	17/15
frequency	17/12	instrument	44/33	container	3/3
time at	31/26	state	11/7	topic	54/43
time through	6/6	property	52/45	measure	31/30
direction	8/8	possessor	43/30	equality	17/5
location	7/5	part	15/8	type	16/12
location at	24/22	location from	28/19	material	44/29

6.3.1 MBL

We have used our own implementation of the nearest-neighbour algorithm on which MBL systems rely on for classification. Each (modifier,noun) pair is represented as a record with 8 attributes:

$$[root_{mod}, POS_{mod}, src_{mod}, WNsense_{mod}, root_{head}, POS_{head}, src_{head}, WNsense_{head}]$$

$root_w$ is the root of the word w, POS_w is the part-of-speech of w, src_w is the word w is derived from, $WNsense_w$ is the *WordNet* sense of w

To a *(noun,modifier)* pair under analysis, we will assign the semantic relation of the closest tagged pair, according to the distance metric used.

The formula to compute the distance between examples i and j is:

$$Dist(i,j) = \sum_k d(a_{ik}, a_{jk})$$

$$d(a_{ik}, a_{jk}) = \begin{cases} WNd(a_{ik}, a_{jk}) & : \quad a_{ik} = root_{mod}, a_{jk} = root_{head} \\ 0 & : \quad a_{ik} = a_{jk} \\ 1 & : \quad a_{ik} \neq a_{jk} \end{cases}$$

where the *WordNet* distance $WNd(w_1, w_2)$ is the length of the path that connects the two synsets to which the words belong, following hypernym/hyponym links.

The downside of this tool, as far as this experiment is concerned, is the fact that each semantic relation has a different "signature" – different attributes from the ones that describe our data are more significant for different semantic relation. The nearest neighbour process on which MBL is based treats all attributes uniformly for all semantic relations. Because of these reasons, we have found that MBL is not a sophisticated enough method to find generalizations in *WordNet* or *Roget's Thesaurus*.

6.3.2 C5.0

We have used the C5.0 implementation of decision tree learning (RuleQuest, 2000). It is a readily available tool much used in the machine learning community. We have used its output in rule format. Although the results obtained were promising, C5.0 did not fit the purpose of the experiment.

In learning the categorization of examples under N classes, each of these N classes will be characterized by rules. In the present experiment, the class is the semantic relation. Each word in an example (noun-modifier pair) is described by the following attributes: In *Roget's Thesaurus*, for the adjective *parental*:

parental	the word w;
a	part of speech of w;
denominal-adj	information about the source of the word (deverbal/ denominal/true adjective or adverb, deverbal/true noun);
parent	the word w_p to which w pertains (or is derived from), according to *WordNet*. If there is no such word, then $w_p = w$;
n	part of speech of w_p;
parentage	first word in the paragraph that best fits w_p's sense;
parentage	headword;
causation	section;
abstract relations	class.

In *WordNet* for the same adjective:

parental,
a,
denominal-adj,
parent,
n,
genitor,
progenitor primogenitor,
ancestor ascendant ascendent antecedent,
relative relation,
person individual someone somebody mortal human soul,
life-form organism being living-thing,
entity something

parental,
a,
denominal-adj,
parent,
n,
genitor,
progenitor primogenitor,
ancestor ascendant ascendent antecedent,
relative relation,
person individual someone somebody mortal human soul,
causal-agent cause causal-agency,
entity something

We have two vectors because the noun *parent* to which *parental* pertains is a hyponym of the noun *person*, which has two hypernym sets:

{ *life form, organism, being, living thing*} and
{ *causal agent, cause, causal agency*}.

Both these vectors will be used in learning.

In a first attempt we have tried to learn characterizations for all classes (semantic relations) at the same time. The system did not perform well. We have changed the experiment from a 30-class problem into 30 binary problems – each semantic relation becomes the *positive class*, all the others are grouped under the generic *negative class*. This introduces two problems:

1. **C5.0 treats both classes, the positive and the negative, in the same way**. We are interested in finding rules that characterize the positive class – what conditions an example must meet for it to be assigned a particular semantic relations. In other words, we want a characterization of the subspace of the semantic space where all the examples are tagged with a certain semantic relation, not a characterization of the complement of this space.

2. **Imbalance**. C5.0 is sensitive to class imbalance. In 550 examples we have represented 30 classes. When we consider examples in one class as positive, and all the other negative, the data set becomes imbalanced (average 1:30). There is no standard method to balance an imbalanced set (Japkowicz, 2001). In a comparative study, Japkowicz (2000) observes that both downsizing of the majority class and resampling of the minority class may have a positive effect on the outcome of the learning process. It all depends on the problem and the tool used.

We have trained the system using several positive to negative ratios (1,2,5). We have also run experiments with and without misclassification costs to compensate for this imbalance. The best results in training were obtained for a ratio of 1:1 of positive to negative examples. We have decided to build a classifier which is trained using the 1:1 ratio, and test it using the distribution in the original set, for each relation. We have performed 5 fold cross-validation for each relation, if the number of positive examples permitted. The number of cross-validations was dynamically adjusted to adapt to relations with few positive examples.

The algorithm performed followed the steps presented in Figure 6.1.

We show in Table 6.2 some of the relations represented in our data set, the number of

```
set N to the number of folds desired
for each relation R
 {
  extract all examples annotated with R
  extract all examples annotated with other relations and
     assign them to the negative class
  if N > number of positive examples
    then set N to the number of positive examples
  split the positive data set into N folds
  split the negative data set into N folds
  for each fold F in the positive set
   {
    form the test set by concatenating F and it's negative counterpart
    for each negative sets N remaining
     {
       randomly downsample N to the size of the positive fold
     }
    form the training set by concatenating the rest of the
       positive sets and all the downsampled negative sets
   apply C5.0 on the training and testing data
   }
  combine the results obtained for each fold
 }
```

Figure 6.1: Learning NMR assignment with C5.0

cross-validation folds, the mean error, standard variation and standard error[2]. We also show for comparison the baseline of comparison for each relation - the error obtained by classifying everything as negative (because negative is the majority class in the testing data). This baseline varies from relation to relation, because it is actually the ratio of positive to the size of the whole data set.

Because of the high imbalance of the data, the default error is smaller than the error made by the system on all relations. Also, because of the small size of the annotated data, the errors vary during the experiments quite a bit, as shown by the standard deviation. This type of analysis is not the most appropriate for the data we have. Instead of looking at accuracy, which we know will be low because of the sparsity of the data, we will look at the rules produced, to check whether the system has found interesting generalizations in the ontologies that we use in representing the data.

The rule sets obtained with C5.0 contained rules for both the positive (semantic relation) and the negative class. In order to obtain only the characterization of the positive

[2] A complete table is presented in Appendix B.1

Relation	Baseline error	Nr. of folds	Mean error	Standard deviation	Standard error
Agent	9.52%	5	6.1%	2.0%	0.9
Cause	2.47%	5	32.9%	16.2%	7.2
Effect	4.82%	5	19.0%	9.9%	4.4
Frequency	2.21%	5	11.9%	14.8%	6.6
Instrument	5.73%	5	42.2%	15.0%	6.7
LocationAt	3.12%	5	36.1%	23.3%	10.4
Material	5.73%	5	19.4%	7.2%	3.2
Measure	4.04%	5	16.1%	14.8%	6.6
Object-Property	1.95%	5	0.0%	0.0%	0.0
Object	5.86%	5	16.9%	18.4%	8.2
Possessor	5.60%	5	9.1%	5.2%	2.3
Product	2.60%	5	16.6%	9.1%	4.1
Property	6.78%	5	21.4%	17.1%	7.6
TimeThrough	0.78%	3	33.8%	36.2%	20.9

Table 6.2: Sample of learning results for C5.0

class, and possibly a lower error in classification, we have tried a rule induction learner, in particular RIPPER.

6.3.3 RIPPER

We have used RIPPER to produce only rules that describe the semantic space where examples tagged with the same semantic relation are located (Cohen, 1995). The rules obtained are quite interesting and capture phenomena that are intuitively correct.

RIPPER was used in experiments which had three parameters - lexical resource (possible values: *WordNet/Roget's*), misclassification costs (possible values: used/not used), nominalization/adjectivalization information (possible values: used/not used). All possible combinations of values for these parameters were tried. We will present a sample of the results. In all cases, misclassification costs only increased the number of rules, without modifying the ones obtained without misclassification costs. This parameter had no influence on the rules presented.

RIPPER produced rules in the following format:

$$Class : -Attr_1 = Value_{Attr1}, ...Attr_N = Value_{AttrN}(NC/NM)$$

where *Class* in our binary classification problems will be the relation that is being analyzed. $Attr_X$ is an attribute that characterizes the data, $Value_{AttrX}$ is one of the possible values of $Attr_X$, NC is the number of examples that the rule classifies correctly, and NM is the number of examples that the rule misclassifies. The names of the attributes indicate their source. For example: *hypernyms_depth_3_head* means the hypernym at depth three (counting down from the most general level), in *WordNet*'s hypernym/hyponym hierarchy, of the head in the base NP. The value of such an attribute is a synset.

Cause and Effect

Best results were obtained with *WordNet*. We present them partially here. Information about the source of the words was used.

Cause – *flu virus* – H is the cause of M (H denotes the head of the noun-phrase, M the modifier).

cause :- hypernyms_depth_3_modifier = {physiological state} (9/2)

Effect – *exam anxiety* – H is the effect of M.

effect :- hypernyms_depth_2_head = {condition, status},
hypernyms_depth_4_head = {ill health, unhealthiness, ... }. (7/1)
effect :- hypernyms_depth_2_head = {happening, occurrence, ... }
head_source = deverbal_noun. (6/1)

It might seem a mistake to have in the same rule several hypernyms of the head word or of the modifier. The structure defined by IS-A links in *WordNet* is a graph. A synset may have several hypernyms and hyponyms. Specifying two hypernyms at different levels in the hierarchy for the same word sense serves as disambiguation among the possible senses (represented as paths in this graph).

Agent

Information about the source of the words improved quite dramatically the precision and quality of the rule set. Considering that syntactic indicators play a major role in the identification of this relation, it is surprising to see a big difference in the performance of the system depending on the lexical resource used – *WordNet* performed much better, and quite well even without word-source information (deverbal/deadjectival noun, gerund, denominal/deverbal adjective). For comparison, we present a sample of the rules built using *WordNet*, without (1) and with (2) word-source information:

Agent – *student protest* – M is the agent of H.

```
1  agent  :-  hypernyms_depth_3_modifier = {person, individual, ...},
              hypernyms_depth_4_modifier = {leader}. (22/1)
   agent  :-  hypernyms_depth_3_modifier = {person, individual, ...},
              hypernyms_depth_1_head = {act, human action, ...}. (18/4)
   agent  :-  hypernyms_depth_3_modifier = {person, individual, ...},
              hypernyms_depth_4_head = {communication}. (8/0)
   agent  :-  hypernyms_depth_2_modifier = {social group},
              hypernyms_depth_1_head = {act, human action, ...}. (6/0)

2  agent  :-  head_source = deverbal_noun,
              hypernyms_depth_3_modifier = {person, individual, ...}. (50/4)
   agent  :-  hypernyms_depth_2_modifier = {social group},
              head_source = deverbal_noun, modifier_pos = noun. (8/0)
```

The information that the head is a deverbal noun seems to subsume the fact that its hypernym is the synset {*act, human action, ...*} , which is the criterion used by Hull and Gomez (1996) in deciding whether a noun is a deverbal noun.

RIPPER has found rule sets that characterize well all **Temporal** relations, especially Frequency (*daily news*). The rules are mostly based on attributes that establish the modifier as a temporal indicator.

6.3.4 FOIL

We have tried to improve the results obtained with RIPPER. At the beginning of this section we have described two ways in which one could account for similarities between noun-modifier pairs.

One of them is discovering, after mapping the elements of our pairs onto a semantic space that they form a cluster. The machine learning tools used until now had the purpose of finding whether this assumption is true, and to find the subspaces that characterize the semantic relations.

The second possibility is that the distance between elements of pairs tagged with the same semantic relation is the same. In order to test this possibility we have used FOIL, a relational learner. It will try to find relations between the attributes that describe the data. In other words, it will try to find regularities in the distance between elements of a pair measured along the dimensions of the semantic space onto which we have projected the data.

For example TYPE (*oak tree* - M is a type of H), EQUATIVE (*composer arranger* - M is also H), PART (*board member* - H is a part of M) and WHOLE (*molecular chain* - M is a part of H), will be better explained by a system that can extract relations between attributes.

On most relations FOIL produces results quite similar to RIPPER. It did not discover the rules to characterize the relations mentioned above. We will explain why, taking TYPE as an example. In this relation, the head noun is a hypernym of the modifier, but not necessarily the first hypernym, as in the following NPs:

nervous system - nervous (sense 3) → (pertains to noun) nervous system → system
oak tree - oak (sense 2) → tree

The system does not perform as expected. A manual analysis shows that the problem lies in the projection of the data onto the semantic space provided by the ontologies used. The projection is not consistent. Attributes of different examples may have the same relative values, but they are represented along different dimensions of the space.

For comparison, we show in Appendix B the rules produced with RIPPER and FOIL for a subset of our relations.

This work on automatically assigning semantic relations to noun-modifier pairs was presented more extensively in (Nastase and Szpakowicz, 2003b).

6.4 To Use or Not to Use Lexical Resources

Working with ontologies and intuitively analyzing the symbolic rules obtained with several machine learning systems has led to an observation about ontologies. They are static. People are able to group concepts into many different categories, according to different rules. Ontologies such as the ones that underlie *WordNet* and *Roget's Thesaurus* capture one instance of this classification process.

A similar problem was analyzed by Barrière and Popowich (2000). They show how to extend a type hierarchy with non-lexical concepts. These concepts are defined by clustering words or word senses in the hierarchy based on similarity criteria other than IS-A relations. For example, *truck, helicopter, airplane*, etc. are all grouped under the head *machine* and *camel* and *donkey* are grouped under the head *animal*. Yet they share something in common that is not captured in the hierarchy - they can all be used for carrying stuff. The hierarchy can be augmented by introducing a node *carrier*, which can be the superclass of all the words that express entities that can be used for carrying.

In order to capture different instances of concept grouping, one can look at the way concepts are used, and how they can be clustered together based on the usage patterns. This can be done through corpus analysis, by extracting collocation information. There are approaches that consider the words as such, whereas other consider word senses, which bring us closer to the concepts that underlie words.

Kilgarriff and Tugwell (2001) present work in building *word sketches* – a characterization of words based on collocation information extracted from the 100-million word British National Corpus. Words are grouped based on some syntactic pattern that they share.

For example, all the words that were found as syntactic *subjects* of the verb *bring* (in the BNC) are clustered (the number represents the number of times the word has appeared as a subject of the verb):

> *waiter 29, effort 104, attempt 87, waitress 10, Yorkshire 30, proceedings 28, zahara 3, slattern 3, move 41, helping 12, wearside 3, servant 20, linda 4, conscription 4, sapt 3, aim 27, porter 7, stork 3, change 92, action 65*

It is obvious that the cluster is not coherent. Based on other collocation information and by cross-referencing, the cluster can be split into smaller, more semantically close, subclusters.

Pantel and Lin (2002) also present work based on clustering of concepts using contextual information from texts. These clusters are used to disambiguate word senses – a word sense will belong to a certain cluster if it shares more of its collocation information with other word senses in that cluster than with any other word senses. The clusters obtained are more semantically compact than the word sketches, but there may be outliers, as in the following cluster, which was built starting from the noun *book*:

> *N1113 grocery store, supermarket, drugstore*[3]

[3] The number next to each phrase represents a measure of how well the word fits into this cluster.

> *grocery store 0.552476, supermarket 0.500276, ... ticket booth 0.16505, newsstand 0.162014, carwash 0.156455, motel 0.150307, ticket office 0.139539, junkyard 0.136729, gambling casino 0.133653, Safeway 0.125724, Costco 0.124233, book 0.122093, driving school 0.120992, automatic teller machine 0.115331, checkout 0.114334, laundry 0.114016, ...*

TANKA's processing starts with very little manually encoded knowledge – three very small dictionaries of relational markers (14 CLR, 302 Case and 48 NMR markers). Heuristics based on case pattern matching are incorporated in the semantic analysis module.

It is worth introducing lexical resources in TANKA only if the performance gain is greater than the overhead in operation.

The overhead is a step of word sense disambiguation (WSD). There are two approaches to WSD:

- **lexical sample**: find all possible senses for the occurrences of a given word in a corpus (Yarowsky, 1995), (Ng and Lee, 1996).

- **all words strategy**: disambiguate all words in a text (Stevenson and Wilks, 2001), (Stetina et al., 1998).

Word sense disambiguation algorithms are not yet close to 100% accuracy on general texts. The results of the SENSEVAL competitions started in 1999 show the evolution of state of the art systems that deal with this issue (Kilgarriff and Palmer, 2000), (Preiss and Yarowsky, 2001). SENSEVAL competitions provide a large corpus and a list of words whose instances in the given texts must be disambiguated. The best system achieves a precision of 75% (80% for nouns and about 70% for verbs, cf. (Kilgarriff and Rosenzweig, 2000)). The SENSEVAL competitions use *WordNet* as the reference for word senses.

A side effect of this word sense disambiguation competition was the discovery of *WordNet*'s weakness, as far as this particular task is concerned – it is too fine grained (Mihalcea and Moldovan, 2001). The distinctions between senses are sometimes too fine, even for human judges.

6.5 Conclusions

The experiments performed at the noun phrase level show that using ontologies and a machine learning system we can find justifications, in terms of generalization in an ontology, about why the same semantic relation describes the interaction between many pairs of concepts. These justifications, in forms of rules, can aid in the automatic assignment of semantic relations.

The resources we used in our experiments had a few shortcomings:

- The resources do not provide all the information that the system needs. We would like derivational information about words to be available.

- The resources are static, and capture only one instance of concept grouping, according to the IS-A relation. We could find few interesting generalizations of concepts involved in the same semantic relations based only on hypernym/hyponym relations.

- Using lexical resources introduces the overhead of performing word sense disambiguation, in which each word in the text is linked to its appropriate entry in the resource.

The use of the lexical information that the resources provide takes us one step closer to automating semantic relation assignment. None of the issues mentioned above presents an insurmountable problem.

Derivational information about words is becoming available. A new version of *WordNet* has just been released. It contains derivational links between the open-class words in its hierarchies. Another resource recently made available is CatVar (CatVar2.0, 2003). Using information from various resources (*WordNet*, *LDOCE*, ...) it groups words that share the same root, although without giving explicit derivational information.

Instead of using fixed ontologies, as the ones underlying *WordNet* and *Roget's Thesaurus*, we have the option of using collocational information extracted from corpora. Kilgarriff and Tugwell (2001) and Lin and Pantel (2002) extract knowledge that captures all aspects of usage of every open-class word in the corpora each of these projects works with. The clusters obtained are not as coherent and error-free as the synsets in *WordNet* and semicolon groups in *Roget's Thesaurus*. It would be interesting nonetheless to find if the trade-off of accuracy versus coverage is worthwhile.

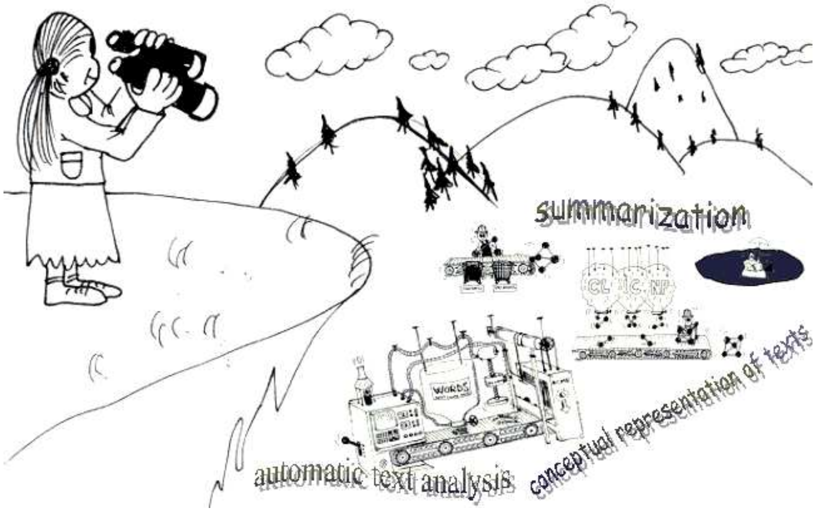
An experiment using a corpus already tagged with word senses would allow us to bypass the word-sense disambiguation task, and concentrate on the contribution that access to an ontology would bring to the task of automatically tagging pairs with semantic relations. Such a resource exists. SemCor consists of a subset of texts from the Brown Corpus, in which all open-class words are tagged with *WordNet* senses.

The experiment described in Chapter 5 does not make use of these resources. The reason was that we wanted to test separately the two aspects of semantic relations that we have investigated and reported in this dissertation. Chapter 5 tests the improvements that a unified view of semantic relations brings to text analysis. This chapter investigated similarities between concepts connected by the same semantic relation through the use of ontologies.

We aim to combine these two aspects into one text analysis and knowledge acquisition system that uses both syntactic and semantic information to pair concepts extracted from texts and assign a semantic relation that describes the interaction between these concepts.

Chapter 7

Future Work



In order to address the problem of semi-automatically assigning semantic relations to pairs of units extracted from texts, we have touched on a number of issues, which for time reasons have remained incompletely explored.

7.1 Towards an Automatic System for Semantic Relation Assignment

We have briefly explored the use of lexical resources for assigning semantic relations to modifier-noun pairs. We have performed experiments with the ontologies underlying the IS-A structures of *WordNet* and *Roget's Thesaurus.* They have shown that ontologies can be used to generalize to classes of concepts between which the same semantic relation holds.

We would like to take this experiment further. We would like to build a system which makes use of ontologies, derivational information about words, and anything that could aid in the automatic assignment of relations. The system will combine the heuristic developed and presented for the current version of HAIKU with information about the word sense extracted from an ontology and with derivational information, and it will use a combined list of relations across levels.

The processing will start using the existing heuristic, and will rely on the user to accumulate some tagged data. After a certain amount of new information is gathered, a machine learning step will be used to extract rules from the tagged data. Further processing will be done using these rules found, and feedback from the user. The representation of the data for machine learning will include all the indicators used in the heuristic, so that by transition towards learning, the information on which the system relied initially can still be used.

The rules on which classification into semantic relations is based will be updated after a certain amount of negative data has gathered against them. A new set of rules will be learned from the existing data, and the processing will continue.

The purpose of this experiment would be to see how much closer to full automation we can take a system that has suitable lexical resources available to it.

It would also be of interest to examine the evolution of the set of rules extracted during processing. We have discussed in Section 6.2 our choice of machine learning tools that produce easily understandable sets of rules. The purpose is to be able to analyze the results obtained, and extract conclusions from the way the rules change during the experiment. We would like to see how general or how specific to a certain domain these rules are, and whether they capture salient information from the domain of the analyzed text.

7.2 A Customizable Knowledge Acquisition System

We have seen through the review of semantic relations that there is no agreement among researchers in NLP on one list of semantic relations to be used for text analysis. This is understandable, since different processing tasks have different requirements: for domain-specific texts one would require a specific list of relations that capture the essence of that domain, whereas for general texts such a list would be useless.

We propose, then, to implement a customizable system, in which an expert can give the system a list of semantic relations that are appropriate for the domain of the text analyzed. An even more flexible system would allow a user to modify this list during processing.

The uniform processing of texts that we have implemented facilitates building such a system. Not only can the user specify the semantic relations that he is interested in, but also the type of syntactic units he wants the system to extract, and the syntactic analyzer that will parse the text.

7.3 A Flexible List of Semantic Relations

We have mentioned briefly at the end of Section 2.3 the possibility that in order to describe the semantic relations between larger units of texts we need relations other than those we propose. It is possible to require a chain of relations, or just new simple relations. Also, when processing texts specific to a certain field, the user may feel that he needs another relation to better describe a specific interaction between concepts. The system should allow him to add a new relation as needed.

In order to address these issues, we envision a system that has a flexible behaviour. It will start with a small list of semantic relations, which would be the core, and the user may add new relations as necessity arises. It would be interesting, in such a setting, to analyze the evolution of the set of semantic relations: how fast it grows, how large it gets, after how much processing it stops growing.

7.4 Beyond the Level of the Sentence

The discussion about surface forms of concepts was restricted to forms that are found inside a sentence. The reason is that the level of the sentence has been analyzed and formalized enough to give us structure and indicators on which we can base our semantic analysis. It would be interesting to extend this discussion to levels above the sentence, and investigate automating the assignment of semantic relations at these levels as well. An interesting aspect of such an investigation would be to test whether the list of relations that we propose for units inside a sentence is appropriate for describing relations between sentences or larger units of text. Sometimes we need to fill the gaps between sentences with background information, as it was the case for the example (15) presented in Section 2.3, which we repeat here:

283 *We were late for the party. The engine broke down.*

In order to connect the two sentences we need to infer that we were going to the party by car whose engine broke down. Part of the missing information comes from common sense knowledge – that engines are inside vehicles, and if engines do not work, vehicles do not move. While we can just assign a Cause relation to describe the interaction between the two sentences, it would be much more interesting and informative if we could reconstruct the chain that connects the two events.

7.5 Below the Level of the Sentence

Semantic relations describe relations between concepts, but these concepts need not be conveyed by two different words, phrases or clauses. A pair of concepts can be conveyed by one word.

There can be several such situations.

One situation is when we have a derived word, like the verb *tape*, for example, which is a denominal verb, whose corresponding noun is *tape*. When we say *I want to tape this movie.*, we mean that we want to use a *tape* in such a way that the movie is recorded on it. So inside the verb *tape* we can perceive a connection between the entity tape and the occurrence taping. We have briefly explored this situation by using dictionary definitions from LDOCE. The results of our experiments are reported in (Nastase and Szpakowicz, 2003a). This experiment may be taken further, and attempt to link all concepts evoked in the definition of a certain concept. This is similar to work done to extract knowledge from dictionaries (Barrière, 1997).

Another situation is when we are analyzing a complex word, formed through agglutination. Some words obtained through this process have acquired a meaning which is far from the meaning of the words it is composed from. For some others, it may be interesting to know the interaction between the concepts behind the words they are formed from. This may be particularly interesting for languages like German, for example, which allow for the formation of new words through agglutination just as easily as English allows for base noun-phrase formation.

7.6 Recovering Implicit Information

Part of the unification of the three separate lists of relations we have started with was the design of a structure for each relation that comprises the necessary and sufficient elements for its expression. Some of these structures were designed to allow for various surface structures. For example, for causality relations, we have established that there are always two occurrences connected through such a semantic relation. However, sometimes in texts only one of the occurrences is expressed as a clause, while the other may surface as a noun phrase. There are also cases when both occurrences surface as noun phrases. The structures were designed so that information about implied predicates can be recovered. The text analyzed contains sentences with a quite simple structure – both from the point of view of syntax and semantics (see Section 5.6). It did not give us the opportunity to make good use of the structures designed. The system should be put to the test on a text with a wider variety of surface expressions for the same concept.

7.7 Text Exploration

By unifying semantic relations across syntactic levels we have taken one step towards a more compact representation of the text, by recognizing different surface forms of concepts. Other steps should be taken. One is to perform reference resolution across the document, to collapse all instances of one concept into one node. Then we would be able to obtain a representation of the text in terms of concepts connected by semantic relations that show how they interact in the given text. We could then explore each concept, analyze its network of relations in order to find all the other concepts in the text with which it interacts, and the nature of these interactions.

7.8 Summarization

We can imagine exploring the use of semantic relations for summarization.

Many text analysis approaches rely on extracting key words or phrases in order to capture the topic of a document. Separate entities, however, do not manage to convey the message in a text. People will usually try to find a way in which they can be connected in a coherent way.

We could extract instead concepts connected by semantic relations. It is possible that one of these concepts is a salient phrase, whereas the other could not be picked by statistical analysis of the text, yet the link between the two captures an interesting aspect of the document.

There can be several ways in which we could make use of semantic relations.

- A number of summarization systems compete annually in a summarization competition. It would be interesting to test whether a combination of these systems would work better than any of them taken separately. In order to combine the summaries they produce, we extract from each the concepts and the semantic relations between them, and we choose from these the ones we find in all, or in most, of the summaries. Chances are that if several systems consider a certain pair of interacting concepts interesting, it will be.

- We could experiment with a system that extracts salient noun phrases from the document using statistical methods, and then follow the text to find *semantic chains* – subsets of our set of phrases, which are connected through semantic relations. We could thus build a graph, and propose it as a summary.

- We could take the previous idea one step further, and compare the graph we have created with author's abstract, to find if and where they overlap. The graph built using key phrases may be larger than the graph built on the abstract. If the graph for the author's abstract corresponds to a subgraph of the graph for key phrases, machine learning techniques can help find what features distinguish this subgraph from the rest of the graph. By extracting parts of an abstract built using key phrases, we could automatically obtain a good abstract.

7.9 Exploring Other Languages

In Section 3.6 we have briefly looked at other languages. We wanted to test whether the phenomena that we have found to relate different forms of the same concept are particular to English. We have observed that they are not. Three of the four major phenomena we propose, *metonymy, equivalence* and *deletion* are more conceptual in nature. If people with different backgrounds share the same cognitive processes, these phenomena could be universal. The other phenomenon, word formation, is language-specific. But all languages have ways of forming new words, or compacting phrases into smaller units, although to a varying degree. It would be an interesting projects to explore such phenomena as well as semantic relations across languages and cultures. They may well be universal.

7.10 Further Down the Road

In order to concentrate on the issue of semantic relations across syntactic levels, we have simplified this problem, by neglecting certain attributes of speech, such as emphasis and focus, and also quantifiers, cardinality, mode, etc. In order to preserve the richness of the message, these must be incorporated back into the analysis, and the subtle differences they introduce must be accounted for.

Chapter 8

Conclusions

Without semantic relations we cannot establish the meaning of a text. They show us how individual pieces connect together to express the idea behind both larger and smaller fragments of text. While they are not physically present in texts, we perceive them through our effort of making sense of what is said.

Native speakers of a language are not conscious of the structural rules and restrictions that they may apply when producing language. On the other hand, linguists, in particular those concerned with syntax, tend to impose on utterances order that formal grammars suggest. We have found this order mirrored in the treatment of semantic relations, by analyzing them depending on the syntactic level they pertain to. Each level has its own structures and complexity. Trying to map semantic relations onto these structures has its own appeal which has captured the interest of many researchers, as we have seen throughout this dissertation. There are also those who have chosen to look at particular relations, without much interest for the grammatical structures that serve to carry them.

We have tried to bring together these views, by letting the grammatical order and distinction fall in the background. Semantic relations hold between concepts. However, it is language that brings forth ideas through words, phrases, sentences, paragraphs, and larger bodies of text. Therefore we have anchored the conceptual view of semantic relations into the structures of language provided by grammar for three syntactic levels: noun-phrase, intra-clause and clause level.

8.1 Goals Revisited

In order to make a connection between the conceptual and the linguistic level, we have looked into how concepts surface in language. We have looked at how people produce speech, both naturally, from the perspective of psycholinguistics, and artificially, from the point of view of text generation systems and grammars.

What we have learned is that the pre-linguistic form of a thought is not conscious. When a thought takes a linguistic form it leaps from the subconscious to the conscious mind, and finding how the thought received its linguistic form is still a research subject. The field of language production has proposed several models, based on researching errors produced in speech. It is believed that such errors are produced because of the subconscious (Freud, 1901), and they would therefore give us a glimpse into the workings of our mind below the conscious level.

Text generation systems and generative grammars postulate some knowledge representation form with which they start their generation process. For text generation the input is represented as a conjunction of predicates, and generative grammar assumes the existence of a deep structure which can be modified by surface rules. The deep structure is innate.

The same idea/message/thought (in psycholinguistics) or deep structure (in generative grammars) can take different surface forms. Using this fact, we have looked at alternative expressions of concepts and we have identified phenomena that account for the same idea (relative to a context) surfacing in different forms. Some of the phenomena discovered pertain to a psychological level (metonymy, equivalence, deletion). Others are grammatical in nature (word-formation). We have looked very briefly at other languages to rule out specificity of these phenomena to English only. We have observed that we can find manifestations of the phenomena discovered in several languages of the Indo-European family (Germanic, Romance and Slavic languages).

Once we have shown that there are psychological and grammatical bases to affirm that semantic relations are the same across syntactic levels, we put this idea to a test. The three lists of relations from which this research started were combined into one, and the resulting list was tested in a text analysis and knowledge acquisition system – TANKA. A previous experiment using this system has served as a baseline for performance comparison. We have modified the semantic analysis module of the system to allow for the use of one list of semantic relations and for the unified treatment of syntactic units extracted from text, regardless of the syntactic level to which they pertain.

The results obtained show that the system's performance is improved with the use of a combined list of relations. Although only the behaviour in assigning relations at the intra-clause level can be compared with the previous system, we observe that the new HAIKU starts to learn earlier. The system starts making good suggestions by the fourth sentence analyzed, and by the 50th sentence it relies more on its own resources than on the user. In the previous version, the system processed about half the input sentences (around 250 sentence) before the number of *accept* or *choose* user actions surpassed the *supply* requests.

We have also explored the use of lexical resources, and augmenting these resources with the type of information that would be relevant for the TANKA system. The experiments performed with lexical resources, in particular *WordNet* and *Roget's Thesaurus*, have shown that the use of ontologies could further improve the learning capabilities of our semi-automatic system. Using such ontologies requires an extra step of word-sense disambiguation, which at this point would introduce more errors in processing. With the development of better word-sense disambiguation systems, resources like this could be incorporated in the semantic analysis module.

8.2 Closing Words

We presented in this dissertation a systematic account of a unified view of semantic relations across syntactic levels, in which we consider meaning in terms of concepts and links between them. This unified view brings the meaning of a text closer to the essence of what a speaker wants to convey. We have concentrated on the big picture, and left some utterance attributes for future analysis. We should, for example, study the effect on semantic relations of co-reference, focus and quantification.

References

Thomas Ahlswede and Martha Palmer. 1988. Parsing vs. text processing in the analysis of dictionary definitions. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL 88)*, pages 217 – 224, Buffalo, NY.

James F. Allen. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154.

John R. Anderson and Gordon H. Bower. 1973. *Human Associative Memory.* Winston, Washington, D.C.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 98)*, pages 86–90, Montreal, Canada.

Ken Barker and Sylvain Delisle. 1996. Experimental validation of a semi-automatic text analyzer. Technical Report TR-96-01, Department of Computer Science, University of Ottawa.

Chris Barker and David Dowty. 1993. Non-verbal thematic proto-roles. In Amy Schafer, editor, *NELS 23*, volume 1, pages 49–62, Amherst. GSLA.

Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun-modifier relationships. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 98)*, pages 96–102, Montreal, Canada.

Ken Barker, Terry Copeck, Sylvain Delisle, and Stan Szpakowicz. 1997a. Systematic construction of a versatile case system. *Journal of Natural Language Engineering*, 3(4):279–315.

Ken Barker, Sylvain Delisle, and Stan Szpakowicz. 1997b. Test-driving TANKA: Evaluating a semi-automatic system of text analysis for knowledge acquisition. In *Proceedings of the 12th Canadian Conference on Artificial Intelligence (CAI 97)*, pages 60–71, Vancouver, BC, Canada.

Ken Barker. 1998. *Semi-Automatic Recognition of Semantic Relationships in English Technical Texts.* Ph.D. thesis, University of Ottawa, Department of Computer Science. http://www.cs.utexas.edu/users/kbarker/thesis.

Caroline Barrière and Fred Popowich. 2000. Expanding the type hierarchy with nonlexical concepts. In *Proceedings of the 13th Conference of the Canadian Society for Computational Studies of Intelligence (CAI 2000)*, pages 53–68, Montreal, Quebec, Canada.

Caroline Barrière. 1997. *From a Children's First Dictionary to a Lexical Knowledge Base of Conceptual Graphs.* Ph.D. thesis, Simon Fraser University, Alberta, Canada.

Caroline Barrière. 2001. Investigating the causal relation in informative texts. *Terminology*, 7(2):135–154.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the Conference on Human Language Technologies (HLT-NAACL 2003)*, pages 16–23, Edmonton, Alberta, Canada.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL 2001)*, pages 50–57, Toulouse, France.

John A. Bateman, Elisabeth A. Maier, Elke Teich, and Leo Wanner. 1991. Towards an architecture for situated text generation. In *Proceedings of the International Conference on Current Issues in Computational Linguistics (ICCICL 91)*, pages 336–349, Penang, Malaysia.

Kathryn Bock and Willem Levelt. 1994. Language production: Grammatical encoding. In Gernsbacher (Gernsbacher, 1994), chapter 29, pages 945–984.

Dwight Bolinger. 1967. Adjectives in English: Attribution and predication. *Lingua*, 18:1–34.

Dwight Bolinger. 1977. *Meaning and Form*. Longman, London and New York.

Ted Briscoe. 1991. Lexical issues in natural language processing. In E. Klein and F. Veltman, editors, *Natural Language and Speech*. Springer-Verlag.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of 5 measures. In *NAACL 01: Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 29–34, Pittsburg, PA, USA.

Brian Butterworth. 1980a. Introduction: A brief review of methods of studying language production. In *Language Production (Vol.1): Speech and Talk* (Butterworth, 1980b), pages 1–17.

Brian Butterworth, editor. 1980b. *Language Production (Vol.1): Speech and Talk*. Academic Press, London, UK.

Brian Butterworth. 1980c. Some constraints on models of language production. In *Language Production (Vol.1): Speech and Talk* (Butterworth, 1980b), pages 423–459.

George Cardona. 1976. *Panini: A Survey of Research*. Mouton, The Hague.

Jason Catlett. 1991. Megainduction: a test flight. In *Proceedings of the 8th International Workshop on Machine Learning*, pages 569–599, Ithaca, NY.

CatVar2.0. 2003. The Categorial Variation Database (English). http://clipdemos.umiacs.umd.edu/catvar.

Martin Chodorow, Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (ACL 85)*, pages 299–304, Chicago, IL.

Noam Chomsky. 1965. *Syntactic Structures.* Mouton, The Hague.

Noam Chomsky. 1966. *Topics in the Theory of Generative Grammar.* Mouton, The Hague.

Noam Chomsky. 1970. Remarks on nominalizations. In Roderick Jacobs and Peter Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 184–221. Ginn and Co., Waltham, MA, USA.

Noam Chomsky. 1982. *Lectures on Government and Binding.* Foris, Dordrecht.

Noam Chomsky. 1995. *The Minimalist Program.* MIT Press, Cambridge, MA.

Morten H. Christiansen, Nick Chater, and Mark S. Seidenberg, editors. 1999. *Cognitive Science (special issue): Connectionist Models of Human Language Processing: Progress and Prospects*, volume 23(4).

Peter Clark and Bruce Porter. 1997. Building concept reprezentations from reusable components. In *Proceedings of the 14th Meeting of American Association for Artificial Intelligence (AAAI 97)*, pages 367–376, Providence, Rhode Island.

Stephen Clark and David Weir. 2001. Class based probability estimation using a semantic hierarchy. In *Proceedings of the 2nd Meeting of the North American chapter of the ACL (NAACL 2001)*, pages 95–102, Pittsburg, PA, USA.

William Cohen. 1993. Efficient pruning methods for separate-and-conquer rule learning systems. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 93)*, pages 988–994, Chambery, France.

William Cohen. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, pages 115–123, Lake Tahoe, California.

Terry Copeck, Sylvain Delisle, and Stan Szpakowicz. 1992. Parsing and Case Analysis in TANKA. Technical report, Computer Science Department, University of Ottawa.

Terry Copeck, Ken Barker, Sylvain Delisle, Stan Szpakowicz, and Jean-François Delannoy. 1997. What is technical text? *Language Sciences*, 19(4):391–424.

Thomas Cover and Peter Hart. 1967. Nearest neighbour pattern classification. *IEEE Transactions on Information Theory*, 13:21–27.

Alan D. Cruse. 1973. Some thoughts on agentivity. *Journal of Linguistics*, 9:1–204.

Walter Daelemans, Antal van den Bosch, and Jakub Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34(1):11–34.

Donald Davidson. 1967. Causal relations. In *Essays on Actions and Events*, pages 149–162. Clarendon Press, Oxford. second edition.

Donald Davidson. 1984. *Essays on Truth and Interpretation.* Oxford University Press, Oxford.

Ferdinand de Saussure. 1959. *Course in General Linguistics.* Philosophical Library, New York. Edited by Charles Bally and Albert Sechehaye. Translated from French by Wade Baskin.

Sylvain Delisle, Terry Copeck, Stan Szpakowicz, and Ken Barker. 1993. Pattern matching for case analysis: A computational definition of closeness. In *Proceedings of the 5th International Conference on Computing and Information (ICCI-93)*, pages 310–315, Sudbury, ON, Canada.

Sylvain Delisle. 1994. *Text Processing Without a-priori Domain Knowledge: Semi-Automatic Linguistic Analysis for Incremental Knowledge Acquisition.* Ph.D. thesis, University of Ottawa, Department of Computer Science. TR-94-02.

Gary S. Dell, Franklin Chang, and Zenzi M. Griffin. 1999. Connectionist models of language production: Lexical access and grammatical encoding. In Christiansen et al. (Christiansen et al., 1999), pages 517–542.

Gary S. Dell. 1986. A spreading activation theory of retrieval in sentence production. *Psychological Review*, 82:407–428.

Teus A. Van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension.* Academic Press, New York.

Rene Dirven and Marjolijn Verspoor. 1998. *Cognitive Exploration of Language and Linguistics.* John Benjamins Publishing Company, Amsterdam/Philadelphia.

David Dowty. 1979. *Word Meaning and Montague Grammar.* Reidel, Dordrecht.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67:547–619.

David Dowty. 2002. 'The Garden Swarms with Bees' and the Fallacy of 'Argument Alternation'. In Yael Ravin and Claudia Leacock, editors, *Polysemy: Theoretical and Computational Approaches*, pages 111–128. Oxford University Press, Oxford.

Mark Dras. 1997. Reluctant Paraphrase: textual restructuring under an optimisation model. In *Proceedings of the 5th bianual meeting of the Pacific Association for Computational Linguistics (PACLING 97)*, pages 98–104, Ohme, Japan.

Lexicon Interest Group EAGLES. 1998. Linguistic aspects of lexical semantics.

Michael Elhadad, Khatleen McKeown, and Jaques Robin. 1997. Floating constraints in lexical choice. *Computational Linguistics*, 23(2):195–239.

Michael Elhadad. 1995. Using argumetation in text generation. *Journal of Pragmatics*, 24:189–220.

James Fan, Ken Barker, Bruce Porter, and Peter Clark. 2001. Representing roles and purpose. In *Proceedings of the 1st International Conference on Knowledge Capture*, pages 38–43.

Gilles Fauconnier. 1985. *Mental Spaces: Aspects of Meaning Construction in Language.* MIT Press, Cambridge, MA.

Charles Fillmore and Beryl T. Atkins. 1998. FrameNet and lexicographic relevance. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Spain.

Charles Fillmore. 1968. The case for case. In Emmond Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston.

Charles Fillmore. 1977. The case for case reopened. *Syntax and Semantics 8: Grammatical Relations*, pages 59–81.

J. Fodor. 1975. *The Language of Thought.* Thomas Crowell, New York.

Sigmund Freud. 1901. *Psychopathology of Everyday Life.* T. Fisher Unwin, London. translation by A. A. Brill (1914) http://psyclassics.yorku.ca/Freud/Psycho/index.htm.

Victoria A. Fromkin and Nan Bernstein Ratner. 1998. Speech production. In Jean Berko Gleason and Nan Bernstein Ratner, editors, *Psycholinguistics*, chapter 7, pages 309–346. Harcourt Brace College Publishers, San Diego, CA, USA.

Victoria A. Fromkin. 1971. The nonanomalous nature of anomalous utterances. *Language*, 47:27–52.

Merrill Garrett. 1980. Levels of processing in sentence production. In Butterworth (Butterworth, 1980b), pages 177–220.

Merrill Garrett. 1984. The organization of processing structure for language production: Applications to aphasic speech. In D. Caplan, A. R. Lecours, and A. Smith, editors, *Biological Perspectives on Language*, pages 172–193. MIT Press, Cambridge, MA, USA.

Morton Ann Gernsbacher. 1991. Comprehending conceptual anaphors. *Language and Cognitive Processes*, 6:81–105.

Morton Ann Gernsbacher, editor. 1994. *Handbook of Psycholinguistics.* Academic Press, New York.

Marcus Giaquinto. 1996. Non-analytic conceptual knowledge. *Mind*, 105:249–268.

Raymond W. Gibbs. 1994. Figurative thought and figurative language. In Gernsbacher (Gernsbacher, 1994), chapter 12, pages 411–446.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Talmy Givon. 1975. Cause and control: On the semantics of interpersonal manipulation. *Syntax and Semantics*, 4:59–89.

Fernando Gomez. 1998a. Linking WordNet verb classes to semantic interpretation. In Sanda Harabagiu, editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 58–64. Association for Computational Linguistics, Somerset, New Jersey, USA.

Fernando Gomez. 1998b. A representation of complex events and processes for the acquisition of knowledge from text. *Kowledge-Based Systems*, 10(4):237–251.

George W. Grace. 1987. *The Linguistic Construction of Reality.* Croom Helm, London; New York.

Jeffrey Gruber. 1965. *Studies in Lexical Relations.* Ph.D. thesis, MIT, Cambridge, MA. Reprinted in Jeffrey Gruber. 1976. *Lexical Structures in Syntax and Semantics.* Part I. North-Holland Publishing Company, Amsterdam.

Patrick Hanks, editor. 1986. *Collins Dictionary of the English Language.* Collins, London and Glasgow.

Zellig Harris. 1970. *Papers in Structural and Transformational Linguistics. Structural and Transformational Linguistics.* Reidel, Dordrecht.

Eduard Hovy. 1988. Planning coherent multisentential text. In *Proceedings of the 26th Meeting of the Association for Computational Linguistics (ACL 88)*, pages 179–188, SUNY, Buffalo, NY.

Eduard Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence: Special Issue on Natural Language Processing*, 63:341–385.

Eduard Hovy. 1996. Chapter 4.1: Language generation overview. In Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology.* http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.

Richard D. Hull and Fernando Gomez. 1996. Semantic interpretation of nominalizations. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 1062–1068, Portland, Oregon, USA.

Nancy Ide and Jean Veronis. 1994. Knowledge extraction from machine-readable dictionaries: An evaluation. In P. Steffens, editor, *Machine Translation and the lexicon*, pages 19–34. Springer-Verlag.

K. A. Subramania Iyer. 1969. *Bhartrhari. A Study of Vakyapadiya in the Light of Ancient Commentaries*. Deccan College Postgraduate Research Institute, Poona.

Ray Jackendoff and David Aaron. 1991. Review of Lakoff and Turner (1989). *Language*, 67(2):320–338.

Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. The MIT Press, Cambridge, MA.

Ray Jackendoff. 1976. Toward an explanatory semantic representation. *Linguistic Inquiry*, 7:85–150.

Ray Jackendoff. 1983. *Semantics and Cognition*. The MIT Press, Cambridge, MA.

Ray Jackendoff. 1989. What is a concept, that a person may grasp it? *Language*, 4:68–102.

Ray Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, MA.

Ray Jackendoff. 1994. *Patterns in the Mind: Language and Human Nature*. Basic Books, New York.

Nathalie Japkowicz. 2000. The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence (IC-AI 2000)*, pages 111–117.

Nathalie Japkowicz. 2001. Concept-learning in the presence of between-class and within-class imbalances. In *Proceedings for the 14th Conference of the Canadian Society for Computational Studies of Intelligence (CAI 2001)*, pages 67–77.

Mark Johnson. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press, Chicago, IL.

George G. Joseph. 1991. *The Crest of the Peacock: non-European Roots of Mathematics*. I.B. Tauris, London, New York.

Jerrold J. Katz and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39:170–210.

Jerrold Katz. 2002. On the general character of semantic theory. In Margolis and Laurence (Margolis and Laurence, 2002), pages Chapter 4:125–150.

Michael H. Kelly. 1998. Rule and idiosyncratically derived denominal verbs: Effects on language production and comprehension. *Memory and Cognition*, 26:369–381.

Adam Kilgarriff and Martha Palmer, editors. 2000. *Computer and the Humanities. Special issue: SENSEVAL. Evaluating Word Sense Disambiguation programs*, volume 34.

Adam Kilgarriff and Joseph Rosenzweig. 2000. English SENSEVAL: Report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1234–1244, Athens, Greece.

Adam Kilgarriff and David Tugwell. 2001. WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Workshop on Collocation: Computational Extraction, Analysis and Exploitation, 39th ACL & 10th EACL*, pages 32–38, Toulouse, France.

Walter Kintsch. 1974. *The Representation of Meaning in Memory.* Erlbaum, Hillsdale, NJ.

Betty Kirkpatrick, editor. 1987. *Penguin Authorized Roget's Thesaurus of English Words and Phrases.* Penguin books.

Judith Klavans, Martin Chodorow, and Nina Wacholder. 1992. Building a knowledge base from parsed definitions. In George Heidorn, Karen Jensen, and Steve Richardson, editors, *Natural Language Processing: PLNLP Approach.* Kluwer, New York.

Saul Kripke. 1980. *Naming and Necessity.* Harvard University Press, Cambridge, MA.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by.* University of Chicago Press, Chicago.

George Lakoff and Mark Turner. 1989. *More than Cool Reason: A Field Guide to Poetic Metaphor.* University of Chicago Press, Chicago, IL.

George Lakoff. 1970. *Irregularity in Syntax.* Holt, Rinehart and Winston, New York.

George Lakoff. 1987. *Women, Fire and Dangerous Things.* Chicago University Press, Chicago.

Nancy Larrick. 1961. *Junior Science Book of Rain, Hail, Sleet and Snow.* Garrard Publishing Company, Champain,IL.

Richard K. Larson. 1998. Events and modification in nominals. In *Proceedings from Semantics and Linguistic Theory (SALT) VIII*, Cornell University, Ithaca, NY, USA.

Mark Lauer. 1992. Extracting knowledge from machine readable dictionaries. In *Proceedings of the 1st Australian Workshop on NLP and IR*, Melbourne, Australia.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds.* Ph.D. thesis, Department of Computing, Macquarie University, Australia, December.

Eric Laurence and Stephen Margolis. 2002. Concepts and cognitive science. In Margolis and Laurence (Margolis and Laurence, 2002), pages Chapter 1:3–81.

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Journal of Communications of the ACM*, 38(11):33–38.

Rosemany Leonard. 1984. *The Interpretation of English Noun Sequences.* North Holland, Amsterdam.

Willem Levelt. 1989. *Speaking: From intention to articulation.* MIT Press, Cambridge, MA.

Judith N. Levi. 1978. *The Syntax and Semantics of Complex Nominals.* Academic Press, New York.

Beth Levin and Malka Rappaport-Hovav. 1996. From lexical semantics to argument realization. Cambridge Research Surveys in Linguistic series. http://www.stanford.edu/ bclevin/borer96-12-2up.ps.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Dekang Lin and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 577–583, Taipei, Taiwan.

Eric Lormand. 1996. How to be a meaning holist. *Journal of Philosophy*, 93:51–73.

John Lyons. 1995. *Linguistic Semantics: An Introduction.* Cambridge University Press.

Catherine Macleod, Ralph Grishman, Adam Myers, Leslie Barrett, and Ruth Reeves. 1998. NOMLEX: A lexicon of nominalizations. In *EURALEX'98*, Liege, Belgium.

Franson D. Manjali. 1997. Dynamic models in semiotics/semantics. Cyber Semiotic Institute. http://www.chass.utoronto.ca/epc/srb/cyber/manout.html.

William C. Mann and Christian M. I. M. Matthiessen. 1985. Nigel: A systemic grammar for text generation. In R. Benson and J. Greaves, editors, *Systemic Perspectives on Discourse: Selected Papers from the 9th Systemics Workshop.* Ablex, London.

William C. Mann and Sandra A. Thompson. 1986a. Relational propositions in discourse. *Discourse Processes*, 9(1):57–90.

William C. Mann and Sandra A. Thompson. 1986b. Rhetorical structure theory: Description and construction of text structures. In Gerard Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics.* Martinus Nijhoff Publishers, Dordrecht.

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

Eric Margolis and Stephen Laurence, editors. 2002. *Concepts: core readings.* The MIT Press, Cambridge, Massachussets, USA.

Kathleen McKeown, Shimei Pan, James Shaw, Desmond Jordon, and Barry Allen. 1997. Language generation for multimedia healthcare briefings. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP 97)*, pages 277–282, Washington D.C., PA, USA.

Kathleen R. McKeown. 1985. *Text generation: Using discourse strategies and focus constraints to generate natural language text.* Cambridge University Press, Cambridge, UK.

Marie Meteer, David D. McDonald, S. Anderson, D. Foster, L. Gay, A. Huettner, and P. Sibun. 1987. MUMBLE-86: Design and implementation. Technical Report COINS-87-87, University of Massachussets at Amherst.

Adam Meyers, Catherine Macleod, Roman Yangarber, Ralph Grishman, Leslie Barrett, and Ruth Reeves. 1998. Using NOMLEX to produce nominalization patterns for information extraction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 98)*, Montreal, Canada. Workshop: The Computational Treatment of Nominals.

R. Mihalcea and D. Moldovan. 2001. Automatic generation of a coarse grained wordnet. In *NAACL 01: Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 35–41, Pittsburg, PA, USA.

George A. Miller and Philip N. Johnson-Laird. 1976. *Language and Perception.* The Belknap Press of Harvard University Press, Cambridge, Massachussetts.

George A. Miller, Christiane Fellbaum, Derek Gross, Katherine Miller, Richard Beckwith, and Randee Tengi. 1995. *Five Papers on WordNet.* Princeton.

Vidya Niwas Misra. 1966. *The Descriptive Technique of Panini.* Mouton, The Hague.

Tom M. Mitchell. 1997. *Machine Learning.* WBC/McGraw-Hill.

Johanna D. Moore and Swartout William R. 1989. A reactive approach to explanation. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI 89)*, pages 1504–1510, Detroit, MI, USA.

Stephen Muggleton and Luc De Raedt. 1994. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19/20:629–679.

Vivi Nastase and Stan Szpakowicz. 2001. Word sense disambiguation in Roget's Thesaurus using WordNet. In *NAACL 01: Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 17–22, Pittsburg, PA, USA.

Vivi Nastase and Stan Szpakowicz. 2003a. Augmenting WordNet's structure using LDOCE. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2003)*, pages 281–194, Mexico City, Mexico.

Vivi Nastase and Stan Szpakowicz. 2003b. Exploring noun-modifier semantic relations. In *5th International Workshop on Computational Semantics*, pages 285–301, Tilburg, The Netherlands.

Vivi Nastase. 2001. Preparing data for learning noun-modifier semantic relations in base noun phrases, TR-2001-05. Technical report, SITE, University of Ottawa.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL 96)*, pages 40–47, Santa Cruz.

Donald A. Norman and David E. Rumelhart. 1975. *Exploration in Cognition.* Freeman, San Francisco.

Geoffrey Nunberg. 1978. *The Pragmatics of Reference.* Indiana University Linguistics Club, Bloomington, Indiana.

Geoffrey Nunberg. 1995. Transfers of meaning. *Journal of Semantics*, 12(1):109–132.

Giulia Pagallo and David Haussler. 1990. Boolean feature discovery in empirical learning. *Journal of Machine Learning*, 5(1):71–99.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 613–619, Edmonton, Canada.

Cécile L. Paris. 1993. *The Use of Explicit Models in Text Generation.* Francis Pinter, London.

Philip L. Peterson. 1985. *Six Grammatical Hypotheses on Actions, Causes and 'Causes'.* Indiana University Linguistics Club, Bloomington, Indiana.

Steven Pinker. 1995. *The Language Instinct.* Harper Perennial, New York.

Steven Pinker. 1999. *Words and Rules: The Ingredients of Language.* Basic Books, New York.

Judita Preiss and David Yarowsky, editors. 2001. *Proceedings of SENSEVAL-2, Association for Computational Linguistics Workshop*, Toulouse, France.

Paul Procter, editor. 1978. *Longman Dictionary of Contemporary English.* Longman Group Ltd., Essex, UK.

James Pustejovsky. 1995. *The Generative Lexicon.* MIT Press, Cambridge, MA.

James Pustejovsky. 1998. The semantics of lexical underspecification. *Folia Linguistica.*

Hillary Putnam. 1975. The meaning of meaning. In K. Gunderson, editor, *Language, Mind and Knowledge.* University of Minnesota Press, Minneapolis.

J. Ross Quinlan. 1987. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27:221–234.

John Ross Quinlan. 1990. Learning logical definitions from relations. *Machine Learning*, 5(3):239–266.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London and New York.

Andrew Radford. 1997. *Syntactic theory and the structure of English. A minimalist approach*. Cambridge University Press, Cambridge.

Malka Rappaport and Beth Levin. 1988. What to do with theta roles. In W. Wilkins, editor, *Syntax and Semantics (21): Thematic Relations*, pages 7–36. Academic Press, New York.

Malka Rappaport-Hovav and Beth Levin. 1992. -er nominals: Implications for a theory of argument structure. *Syntax and Semantics: Syntax and the lexicon*, 26:127–153.

Victor Raskin and Sergei Nirenburg. 1995. Lexical semantics of adjectives: A microtheory of adjectival meaning. Memoranda in Computer and Cognitive Science MCCS-95-288, Computing Research Lab, New Mexico State University.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems in natural language. *Journal of Artificial Intelligence*, 11:95–130.

Stephen Richardson, William Dolan, and Lucy Vanderwende. 1998. MindNet: Acquiring and structuring semantic information from text. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 98)*, pages 1098–1102, Montreal, Canada.

German Rigau, Horacio Rodriguez, and Eneko Agirre. 1998. Building accurate semantic taxonomies from monolingual mrds. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 98)*, pages 1103–1109, Montreal, Canada.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun-compounds via a domain specific hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 82–90, Pittsburg, PA, USA.

Barbara Rosario, Marti Hearst, and Charles Fillmore. 2002. The descent of hierarchy and selection in relational semantics. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, PA, USA.

Research RuleQuest. 2000. Data mining tools: C5.0 tutorial.
http://www.rulequest.com.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In James L. McClelland, David E. Rumelhart, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume Vol 2: Psychological and Biological Models, pages 216–271. MIT Press, Cambridge, MA.

Elizabeth Scarlett. 2000. An evaluation of a rule-based parser of English sentences. Master's thesis, University of Ottawa, Ottawa, ON, Canada.

Roger C. Schank and R. Abelson. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures.* Erlbaum, Hillsdale, NJ.

Roger C. Schank. 1973. Idemtification of conceptualizations underlying natural language. In R.C. Schank and K.M. Colby, editors, *Computer Models of Thought and Language.* W.H. Freeman, San Francisco, CA.

Roger C. Schank. 1975. *Conceptual Information.* North-Holland Publishing Company, Amsterdam.

Helmut Scharfe. 1977. Grammatical literature. In Jan Gonda, editor, *History of Indian Literature*, volume V, fasc.2. Otto Harrassowitz, Wiesbaden.

Donia R. Scott and Clarisse Sieckenius de Souza. 1990. Getting the message across in RST-based text generation. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation.* Academic Press, New York.

Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic Paraphrase Acquisition from News Articles. In *Proceedings of Human Language Technology Conference*, San Diego, USA.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge Acquisition from the General Public. In *Proceedings of the 1st International Conference on Ontologies, Databases and Application of Semantics for Large Scale Information Systems*, University of California, Irvine, CA.

Harold L. Somers. 1987. *Valency and Case in Computational Linguistics.* Edinburgh University Press, Edinburgh.

John F. Sowa. 1984. *Knowledge Representation: Logical, Philosophical and Computational Foundations.* Addison-Wesley Publishing Company, Reading, MA, Menlo Park, CA.

Jiri Stetina, Sadao Kurohashi, and Makoto Nagao. 1998. General word sense disambiguation method based on a full sentential context. In *Usage of WordNet in Natural Language Processing, Proceedings of COLING-ACL Workshop*, Montreal, Canada.

Mark Stevenson and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–351.

Beth Sundheim. 1993. *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Baltimore, MD.

Beth Sundheim. 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Baltimore, MD.

Whitney Tabor and Michael K. Tanenhaus. 1999. Dynamical models of sentence processing. In Christiansen et al. (Christiansen et al., 1999), pages 491–515.

Leonard Talmy. 1985. How language structures space. In L.P. Acredolo H. L. Pick Jr., editor, *Spatial Orientation: Theory, Research and Application.* Plenum Press, London and NewYork.

Alfred Tarski. 1944. The semantic conception of truth. *Philosophy and Phenomenological Research*, 4:341–375. Reprinted in (Zabeeh et al., 1974, pages 675-712).

Alfred Tarski. 1983. The concept of truth in formalized languages. In J. Corcoran, editor, *Logic, Semantics, Metamathematics*, page 152278. Hackett Publishing Co., Indianapolis.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale.* C. Klincksieck, Paris.

Sandra A. Thompson and Robert E. Longacre. 1985. Adverbial clauses. In Timothy Shopen, editor, *Language Typology and Syntactic Description (vol.2): Complex Constructions*, chapter 4, pages 171–234. Cambridge University Press, Cambridge.

Robert D. van Valin. 1990. Semantic parameters of split transitivity. *Language*, 66:221–260.

Lucy Vanderwende. 1994. Algorithm for the automatic interpretation of noun sequences. In *Proceedings of the 15th International Conference in Computational Linguistics (COLING 94)*, pages 782–788, Kyoto, Japan.

Sholom Weiss and Nitin Indurkhya. 1993. Optimized rule induction. *IEEE Expert*, 8(6):61–69.

Benjamin Lee Whorf. 1956. A Linguistic Consideration of Thinking in Primitive Communities. *Language, Thought and Reality*, pages 65–86.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 95)*, pages 189–196, Cambridge, MA, 1995.

Farhang Zabeeh, E.D. Klemke, and Arthur Jacobson, editors. 1974. *Readings in Semantics.* University of Illinois Press, Urbana, IL, Chicago and London.

Appendix A

Sample representation of relation instances in attribute-value format

A.1 Causality

- Cause: 1 causes 2. 1 is sufficient to cause 2, and 1 is known to exist. As a consequence, 2 will exist.

  (CL) $\underline{\textit{He went blind}_2}$ *because* $\underline{\textit{the snow was shining sharply}_1}$.

$$\left[\begin{array}{ll} \text{CAUSE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{shine} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{snow} \end{array}\right] \end{array}\right] \\ \text{EFFECT} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{go blind} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{he} \end{array}\right] \end{array}\right] \\ \text{INDICATOR} & \text{because} \end{array}\right]$$

  (IC) $\underline{\textit{He went blind}_2}$ *because of* $\underline{\textit{the snow}_1}$.

$$\left[\begin{array}{ll} \text{CAUSE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{snow} \end{array}\right] \end{array}\right] \\ \text{EFFECT} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{go blind} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{he} \end{array}\right] \end{array}\right] \\ \text{INDICATOR} & \text{because} \end{array}\right]$$

  (NP) $\underline{\textit{snow}_1}$ $\underline{\textit{blindness}_2}$

$$\left[\begin{array}{ll} \text{CAUSE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{snow} \end{array}\right] \end{array}\right] \\ \text{EFFECT} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{blindness} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & - \end{array}\right] \end{array}\right] \\ \text{INDICATOR} & - \end{array}\right]$$

- Effect: 2 is the result of 1. (1 causes 2). 2 is the focus

  (CL) $\underline{\textit{The program issued a command}_1}$ *so* $\underline{\textit{the file printed}_2}$.

$$\begin{bmatrix} \text{CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{issue} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{command} \end{bmatrix} \end{bmatrix} \\ \text{EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{print} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{file} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{so} \end{bmatrix}$$

(IC) *The command*$_1$ *caused* *the printing of the file*$_2$.

$$\begin{bmatrix} \text{CAUSE} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{command} \end{bmatrix} \end{bmatrix} \\ \text{EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{print} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{file} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{cause} \end{bmatrix}$$

(NP) *print*$_2$ *command*$_1$

$$\begin{bmatrix} \text{CAUSE} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{command} \end{bmatrix} \end{bmatrix} \\ \text{EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{print} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & - \end{bmatrix}$$

- PURPOSE: 1 is for 2, but 2 does not necessarily come into being.

  (CL) *She took the medicine*$_1$ *so* *the pain should be relieved*$_2$.

$$\begin{bmatrix} \text{CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{take} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{medicine} \end{bmatrix} \end{bmatrix} \\ \text{DESIRED EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{relieve} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{pain} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{so} \end{bmatrix}$$

(IC) *She took the medicine*$_1$ *for* *pain relief*$_2$.

$$\begin{bmatrix} \text{CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{take} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{medicine} \end{bmatrix} \end{bmatrix} \\ \text{DESIRED EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{relief} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{pain} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{for} \end{bmatrix}$$

(NP) *pain-relief*$_2$ *medicine*$_1$

$$\begin{bmatrix} \text{CAUSE} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{medicine} \end{bmatrix} \end{bmatrix} \\ \text{DESIRED EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{relief} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{pain} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & - \end{bmatrix}$$

- Entailment: 1 entails 2. 1 is not known to exist or not, but if it does then necessarily 2 also exists.

(CL) *If* *students work hard*$_1$*,* *they pass their exams*$_2$*.*

$$\left[\begin{array}{ll}
\text{POSSIBLE CAUSE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{work}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Agent}\\ \text{FILLER} & \text{students}\end{array}\right]\end{array}\right]\\
\text{EFFECT} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{pass}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object}\\ \text{FILLER} & \text{exam}\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & \text{if}
\end{array}\right]$$

(IC) *Hard-working students*$_1$ *pass their exams*$_2$*.*

$$\left[\begin{array}{ll}
\text{POSSIBLE CAUSE} & \left[\begin{array}{ll}\text{VERB/STATE} & -\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & -\\ \text{FILLER} & \text{students}\end{array}\right]\end{array}\right]\\
\text{EFFECT} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{pass}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object}\\ \text{FILLER} & \text{exam}\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & -
\end{array}\right]$$

(NP) *a pass*$_2$ *due to* *hard work*$_1$

$$\left[\begin{array}{ll}
\text{POSSIBLE CAUSE} & \left[\begin{array}{ll}\text{VERB/STATE} & -\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & -\\ \text{FILLER} & \text{work}\end{array}\right]\end{array}\right]\\
\text{EFFECT} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{pass}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & -\\ \text{FILLER} & -\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & \text{due to}
\end{array}\right]$$

- Enablement: 1 enables 2. 1 is necessary but not sufficient to make 2 exist.

(CL) *The printer can print*$_2$ *if* *the paper tray is full*$_1$*.*

$$\left[\begin{array}{ll}
\text{POSSIBLE CAUSE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{is full}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Stative}\\ \text{FILLER} & \text{tray}\end{array}\right]\end{array}\right]\\
\text{DESIRED EFFECT} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{print}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Agent}\\ \text{FILLER} & \text{printer}\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & \text{if}
\end{array}\right]$$

(IC) *The printer can print*$_2$ *from a full paper tray*$_1$*.*

$$\left[\begin{array}{ll}
\text{POSSIBLE CAUSE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{full}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Stative}\\ \text{FILLER} & \text{tray}\end{array}\right]\end{array}\right]\\
\text{DESIRED EFFECT} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{print}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Agent}\\ \text{FILLER} & \text{printer}\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & -
\end{array}\right]$$

(NP) no instance found

- DETRACTION: 1 detracts/opposes 2, but the existence of 1 may not be sufficient to prevent 2 from existing.

  (CL) *They persisted*$_2$ *although* *I warned them*$_1$.

$$\begin{bmatrix} \text{OPPOSING CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{warn} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{I} \end{bmatrix} \end{bmatrix} \\ \text{EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{persist} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{they} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{although} \end{bmatrix}$$

  (IC) *They persisted*$_2$ *despite* *my warning*$_1$.

$$\begin{bmatrix} \text{OPPOSING CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{warning} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{I} \end{bmatrix} \end{bmatrix} \\ \text{EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{persist} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{they} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{despite} \end{bmatrix}$$

  (NP) *persistence*$_2$ *despite* *warnings*$_1$

$$\begin{bmatrix} \text{OPPOSING CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{warning} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \\ \text{EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{persistence} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{despite} \end{bmatrix}$$

- PREVENTION: 1 prevents 2. If 1 is known to exist, then 2 necessarily does not exist.

  (CL) *The service did not work* $_2$ *since* *the hard-disk crashed*$_1$.

$$\begin{bmatrix} \text{OPPOSING CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{crash} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{hard-disk} \end{bmatrix} \end{bmatrix} \\ \text{OPPOSED EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{work} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{service} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{since} \end{bmatrix}$$

  (IC) *The service did not work* $_2$ *because of* *a hard-disk crash*$_1$.

$$\begin{bmatrix} \text{OPPOSING CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{crash} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{hard-disk} \end{bmatrix} \end{bmatrix} \\ \text{OPPOSED EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{work} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{service} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{because} \end{bmatrix}$$

  (NP) *service breakdown*$_2$ *on account of* *a crash*$_1$

$$\begin{bmatrix}
\text{OPPOSING CAUSE} & \begin{bmatrix} \text{VERB/STATE} & \text{crash} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \\
\text{OPPOSED EFFECT} & \begin{bmatrix} \text{VERB/STATE} & \text{breakdown} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{service} \end{bmatrix} \end{bmatrix} \\
\text{INDICATOR} & \text{on account of}
\end{bmatrix}$$

A.2 Temporality

- CO-OCCURRENCE: 1 and 2 occur or exist at the same time. 1 and 2 express unbounded time intervals, they both represent occurrences.

  (CL) *He writes novels*$_1$ *while he listens to music*$_2$.

$$\left[\begin{array}{ll}
\text{TIME1} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{write} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{novel}\end{array}\right]\end{array}\right]\end{array}\right] \\
\text{TIME2} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{listen} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{music}\end{array}\right]\end{array}\right]\end{array}\right] \\
\text{INDICATOR} & \text{while}
\end{array}\right]$$

  (IC) *He writes novels*$_1$ *while listening to music*$_2$.

$$\left[\begin{array}{ll}
\text{TIME1} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{write} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{novel}\end{array}\right]\end{array}\right]\end{array}\right] \\
\text{TIME2} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{listen} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{music}\end{array}\right]\end{array}\right]\end{array}\right] \\
\text{INDICATOR} & \text{while}
\end{array}\right]$$

  (NP) *writing novels*$_1$ *while listening to music*$_2$

$$\left[\begin{array}{ll}
\text{TIME1} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{write} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{novel}\end{array}\right]\end{array}\right]\end{array}\right] \\
\text{TIME2} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{listen} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{music}\end{array}\right]\end{array}\right]\end{array}\right] \\
\text{INDICATOR} & \text{while}
\end{array}\right]$$

- FREQUENCY: 1 occurs every time 2 occurs. 1 is an occurrence, 2 can be an occurrence or a point or time interval that appears several times on the time axis.

  (CL) *We play volleyball*$_1$ *every time he visits*$_2$.

$$\left[\begin{array}{ll}
\text{TIME} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{play} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{volleyball}\end{array}\right]\end{array}\right]\end{array}\right] \\
\text{FREQUENCY} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{visit} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{he}\end{array}\right]\end{array}\right]\end{array}\right] \\
\text{INDICATOR} & \text{every time}
\end{array}\right]$$

  (IC) *We play volleyball*$_1$ *every week*$_2$.

$$\left[\begin{array}{ll} \text{TIME} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{play} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{volleyball} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{FREQUENCY} & \left[\begin{array}{ll} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{week} \end{array}\right] \\ \text{INDICATOR} & \text{every} \end{array}\right]$$

(NP) *weekly*$_2$ *game*$_1$

$$\left[\begin{array}{ll} \text{TIME} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{game} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{FREQUENCY} & \left[\begin{array}{ll} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{weekly} \end{array}\right] \\ \text{INDICATOR} & - \end{array}\right]$$

- PRECEDENCE: 1 occurs or exists (or begins to occur or exist) before 2. Either 1 or 2 is an occurrence, the other can be an occurrence or an explicit temporal expression.

(CL) *I watered the flowers*$_1$ *before* *I left for the holidays*$_2$.

$$\left[\begin{array}{ll} \text{PRECEDENT} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{water} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{flower} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{TIME} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{leave} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Purpose} \\ \text{FILLER} & \text{holidays} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{INDICATOR} & \text{before} \end{array}\right]$$

(IC) *I watered the flowers*$_1$ *before* *leaving for the holidays*$_2$.

$$\left[\begin{array}{ll} \text{PRECEDENT} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{water} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{flower} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{TIME} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{leave} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Purpose} \\ \text{FILLER} & \text{holidays} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{INDICATOR} & \text{before} \end{array}\right]$$

(IC) *watered the flowers*$_1$ *before* *8 o'clock in the morning*$_2$.

$$\left[\begin{array}{ll} \text{PRECEDENT} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{water} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{flower} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{TIME2} & \left[\begin{array}{ll} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{8 o'clock} \end{array}\right] \\ \text{INDICATOR} & \text{before} \end{array}\right]$$

(NP) *watering the flowers*$_1$ *before* *leaving for the holidays*$_2$

$$\begin{bmatrix} \text{PRECEDENT} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{water} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{flower} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{leave} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Purpose} \\ \text{FILLER} & \text{holiday} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{before} \end{bmatrix}$$

- TimeAt: 1 occurs when 2 occurs. Both 1 and 2 can be occurrences or explicit temporal expressions.

  (CL) *He traveled there*$_1$ *when they called him*$_2$.

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{travel} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{he} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME AT} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{call} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{they} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{when} \end{bmatrix}$$

  (IC) *He traveled there*$_1$ *last year*$_2$.

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{travel} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{he} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME AT} & \begin{bmatrix} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{[last] year} \end{bmatrix} \\ \text{INDICATOR} & - \end{bmatrix}$$

  (NP) *winter*$_2$ *travel*$_1$

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{travel} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME AT} & \begin{bmatrix} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{winter} \end{bmatrix} \\ \text{INDICATOR} & - \end{bmatrix}$$

- TimeFrom: 1 began to occur when 2 occurred. 1 is an occurrence, 2 can be an occurrence or point/interval in time, but is considered punctual.

  (CL) *He has been playing well*$_1$ *since we coached him*$_2$.

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{play} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{he} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME FROM} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{coach} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{we} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{since} \end{bmatrix}$$

(IC) *He has been playing well*$_1$ *since January*$_2$.

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{play} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{AGENT} \\ \text{FILLER} & \text{he} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME FROM} & \begin{bmatrix} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{January} \end{bmatrix} \\ \text{INDICATOR} & \text{since} \end{bmatrix}$$

(NP) *playing well*$_1$ *since January*$_2$

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{play} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME FROM} & \begin{bmatrix} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{January} \end{bmatrix} \\ \text{INDICATOR} & \text{since} \end{bmatrix}$$

- TimeThrough: 1 existed while 2 existed. 1 is an occurrence, 2 can be either an occurrence that delimits an interval of time, or an explicit time interval.

  (CL) *The band practices*$_1$ *while other students have lunch*$_2$.

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{practice} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{AGENT} \\ \text{FILLER} & \text{band} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME THROUGH} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{have [lunch]} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{AGENT} \\ \text{FILLER} & \text{students} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{INDICATOR} & \text{while} \end{bmatrix}$$

  (IC) *The band practices*$_1$ *during lunch hour*$_2$.

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{practice} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{AGENT} \\ \text{FILLER} & \text{band} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME THROUGH} & \begin{bmatrix} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{[lunch] hour} \end{bmatrix} \\ \text{INDICATOR} & - \end{bmatrix}$$

  (NP) *lunch-hour*$_2$ *practice*$_1$

$$\begin{bmatrix} \text{TIME} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{practice} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{TIME THROUGH} & \begin{bmatrix} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{lunch-hour} \end{bmatrix} \\ \text{INDICATOR} & - \end{bmatrix}$$

- TimeTo: 1 existed until 2 started to exist or occur. 1 is an occurrence, 2 can be an occurrence that is considered punctual, or a point in time.

  (CL) *They partied*$_1$ *until their mother sent them to bed*$_2$.

$$\left[\begin{array}{ll}
\text{TIME} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{party} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{they} \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{TIME TO} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{send} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{mother} \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{INDICATOR} & \text{until}
\end{array}\right]$$

  (IC) *They partied*$_1$ *until 9 o'clock*$_2$.

$$\left[\begin{array}{ll}
\text{TIME} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{party} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{they} \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{TIME TO} & \left[\begin{array}{ll} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{9 o'clock} \end{array}\right] \\
\text{INDICATOR} & \text{until}
\end{array}\right]$$

  (NP) *party*$_1$ *until dawn*$_2$

$$\left[\begin{array}{ll}
\text{TIME} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{party} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & - \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{TIME TO} & \left[\begin{array}{ll} \text{TYPE} & \text{DEFINITE TIME} \\ \text{FILLER} & \text{dawn} \end{array}\right] \\
\text{INDICATOR} & \text{until}
\end{array}\right]$$

A.3 Spatiality

- Direction: 1 is directed towards 2. 2 is not the final point. The final point is not specified.

  (IC) *Look*$_1$ *inside yourself*$_2$ *for the answer.*

$$\left[\begin{array}{ll}
\text{OCC/ENTITY} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{look} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & - \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{DIRECTION} & \text{inside [yourself]} \\
\text{INDICATOR} & -
\end{array}\right]$$

  (NP) *inward* $_2$ *look*$_1$

$$\left[\begin{array}{ll}
\text{OCC/ENTITY} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{look} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & - \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{DIRECTION} & \text{inward} \\
\text{INDICATOR} & -
\end{array}\right]$$

- LocationFrom: 1 starts at 2.

  (IC) *The capital comes*$_1$ *from foreign countries*$_2$.

$$\left[\begin{array}{ll}
\text{OCC/ENTITY} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{come} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{capital} \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{LOCATION FROM} & \text{foreign countries} \\
\text{INDICATOR} & \text{from}
\end{array}\right]$$

  (NP) *foreign*$_2$ *capital*$_1$

$$\left[\begin{array}{ll}
\text{OCC/ENTITY} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{capital} \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{LOCATION FROM} & \text{foreign} \\
\text{INDICATOR} & -
\end{array}\right]$$

- LocationTo: 2 is the end point of 1.

  (IC) *I went*$_1$ *home*$_2$.

$$\left[\begin{array}{ll}
\text{OCC/ENTITY} & \left[\begin{array}{ll} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{go} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{I} \end{array}\right] \end{array}\right] \end{array}\right] \\
\text{LOCATION FROM} & \text{home} \\
\text{INDICATOR} & -
\end{array}\right]$$

  (NP) *homeward*$_2$ *journey*$_1$

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{journey} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{LOCATION FROM} & \text{homeward} \\ \text{INDICATOR} & - \end{bmatrix}$$

- LocationThrough: 1 occurred through 2. 1 is an occurrence, 2 is a non punctual space.

  (IC) *We traveled*$_1$ *all over Europe*$_2$.

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{travel} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{we} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{LOCATION THROUGH} & \text{Europe} \\ \text{INDICATOR} & \text{all over} \end{bmatrix}$$

  (NP) *travel*$_1$ *all over Europe*$_2$

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{travel} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{LOCATION THROUGH} & \text{Europe} \\ \text{INDICATOR} & \text{all over} \end{bmatrix}$$

- LocationAt: 1 is the location of 2. 2 is an occurrence or an entity, 1 is a point in space (or is considered a point).

  (IC) *My home is*$_2$ *in* *this town* $_1$.

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{be} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{home} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{LOCATION AT} & \text{town} \\ \text{INDICATOR} & \text{in} \end{bmatrix}$$

  (NP) *home*$_2$ *town*$_1$ (also spelled *hometown*)

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{home} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{LOCATION AT} & \text{town} \\ \text{INDICATOR} & - \end{bmatrix}$$

- Located: 1 is located at the point indicated by 2. 1 is an occurrence or an entity, 2 is a punctual space (or is considered punctual).

  (IC) *The storm started*$_1$ *in the desert*$_2$.

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{start} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{storm} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{LOCATED} & \text{desert} \\ \text{INDICATOR} & \text{in} \end{bmatrix}$$

(NP) $\underline{\textit{desert}_2}$ $\underline{\textit{storm}_1}$

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{desert} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{LOCATED} & \text{storm} \\ \text{INDICATOR} & - \end{bmatrix}$$

- Orientation: 1 is oriented like 2.

  (IC) $\underline{\textit{The tree stood}_1}$ $\underline{\textit{erect}_2}$ *despite the heavy ice.*

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & \text{stood} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Stative} \\ \text{FILLER} & \text{tree} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{ORIENTATION} & \text{erect} \\ \text{INDICATOR} & - \end{bmatrix}$$

(NP) $\underline{\textit{erect}_2}$ $\underline{\textit{tree}_1}$

$$\begin{bmatrix} \text{OCC/ENTITY} & \begin{bmatrix} \text{TYPE} & \text{OCCURRENCE} \\ \text{FILLER} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{tree} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{ORIENTATION} & \text{erect} \\ \text{INDICATOR} & - \end{bmatrix}$$

A.4 Conjunctive

- CONJUNCTION: both 1 and 2 occur or exist.

  (CL) *The computer runs applications*$_1$ *and the printer prints documents*$_2$.

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{run}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Agent}\\ \text{FILLER} & \text{computer}\end{array}\right]\end{array}\right]\\
\text{OCCURRENCE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{print}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Agent}\\ \text{FILLER} & \text{printer}\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & \text{and}
\end{array}\right]$$

  (NP) *running*$_1$ *and swimming*$_2$ *(are good for you)*

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{run}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & -\\ \text{FILLER} & -\end{array}\right]\end{array}\right]\\
\text{OCCURRENCE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{swim}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & -\\ \text{FILLER} & -\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & \text{and}
\end{array}\right]$$

- DISJUNCTION: either one or both 1 and 2 occur or exist.

  (CL) *The program may terminate*$_1$ *or it may hang indefinitely*$_2$.

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{terminate}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object}\\ \text{FILLER} & \text{program}\end{array}\right]\end{array}\right]\\
\text{OCCURRENCE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{hang}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & \textsc{Object}\\ \text{FILLER} & \text{program}\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & \text{or}
\end{array}\right]$$

  (NP) *painting*$_1$ *or drawing*$_2$

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{paint}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & -\\ \text{FILLER} & -\end{array}\right]\end{array}\right]\\
\text{OCCURRENCE} & \left[\begin{array}{ll}\text{VERB/STATE} & \text{draw}\\ \text{OCCURRENCE PART} & \left[\begin{array}{ll}\text{TYPE} & -\\ \text{FILLER} & -\end{array}\right]\end{array}\right]\\
\text{INDICATOR} & \text{or}
\end{array}\right]$$

A.5 Participant

- CO-AGENT(accompaniment): 1 is accompanied by 2. 2 is a co-agent.

  (IC) $\underline{\textit{We eat supper}}_1$ *with* $\underline{\textit{my family}}_2$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{eat} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{supper} \end{bmatrix} \end{bmatrix} \\ \text{CO-AGENT} & \text{family} \\ \text{INDICATOR} & \text{with} \end{bmatrix}$$

  (NP) $\underline{\textit{supper}}_1$ *with* $\underline{\textit{my family}}_2$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{supper} \end{bmatrix} \end{bmatrix} \\ \text{CO-AGENT} & \text{family} \\ \text{INDICATOR} & \text{with} \end{bmatrix}$$

- AGENT: 1 performs 2.

  (IC) $\underline{\textit{The students}}_1$ $\underline{\textit{protested}}_2$ *against tuition fee increase.*

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{protest} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \\ \text{AGENT} & \text{student} \\ \text{INDICATOR} & - \end{bmatrix}$$

  (NP) $\underline{\textit{student}}_1$ $\underline{\textit{protest}}_2$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{protest} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \\ \text{AGENT} & \text{student} \\ \text{INDICATOR} & - \end{bmatrix}$$

- BENEFICIARY: 1 benefits from 2.

  (IC) $\underline{\textit{The price discount applies}}_2$ *only for* $\underline{\textit{ students}}_1$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{apply} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{discount} \end{bmatrix} \end{bmatrix} \\ \text{BENEFICIARY} & \text{student} \\ \text{INDICATOR} & \text{for} \end{bmatrix}$$

  (NP) $\underline{\textit{student}}_1$ $\underline{\textit{discount}}_2$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{discount} \end{bmatrix} \end{bmatrix} \\ \text{BENEFICIARY} & \text{student} \\ \text{INDICATOR} & - \end{bmatrix}$$

- EXCLUSION: 2 is excluded from 1, or 1 replaces 2.

  (IC) $\underline{\textit{We cooked rice}}_1$ *instead of* $\underline{\textit{potatoes}_2}$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{cook} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{O}\text{\small BJECT} \\ \text{FILLER} & \text{rice} \end{bmatrix} \end{bmatrix} \\ \text{EXCLUDED} & \text{potato} \\ \text{INDICATOR} & \text{instead} \end{bmatrix}$$

  instead of

  (NP) $\underline{\textit{rice}}_1$ *instead of* $\underline{\textit{potatoes}}_2$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{–} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{–} \\ \text{FILLER} & \text{rice} \end{bmatrix} \end{bmatrix} \\ \text{EXCLUDED} & \text{potatoe} \\ \text{INDICATOR} & \text{instead of} \end{bmatrix}$$

- STATIVE: 1 is in a state of 2.

  (IC) $\underline{\textit{The dog}}_1$ $\underline{\textit{is sleeping}}_2$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{be sleeping} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{–} \\ \text{FILLER} & \text{–} \end{bmatrix} \end{bmatrix} \\ \text{EXPERIENCER} & \text{dog} \\ \text{INDICATOR} & \text{–} \end{bmatrix}$$

  (NP) $\underline{\textit{sleeping}}_2$ $\underline{\textit{dog}}_1$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{sleeping} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{–} \\ \text{FILLER} & \text{–} \end{bmatrix} \end{bmatrix} \\ \text{EXPERIENCER} & \text{dog} \\ \text{INDICATOR} & \text{–} \end{bmatrix}$$

- PROPERTY: 1 has the property 2.

  (IC) $\underline{\textit{The dog}}_1$ $\underline{\textit{is brown}}_2$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{be} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{S}\text{\small TATIVE} \\ \text{FILLER} & \text{dog} \end{bmatrix} \end{bmatrix} \\ \text{PROPERTY} & \text{brown} \\ \text{INDICATOR} & \text{–} \end{bmatrix}$$

  (NP) $\underline{\textit{brown}}_2$ $\underline{\textit{dog}}_1$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{–} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \text{–} \\ \text{FILLER} & \text{dog} \end{bmatrix} \end{bmatrix} \\ \text{PROPERTY} & \text{brown} \\ \text{INDICATOR} & \text{–} \end{bmatrix}$$

- Possessor: 1 has 2.

  (IC) *The man has*$_1$ *a long beard*$_2$.

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{have} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Stative} \\ \text{FILLER} & \text{man} \end{array}\right] \end{array}\right] \\ \text{POSSESSOR} & \text{beard} \\ \text{INDICATOR} & - \end{array}\right]$$

  (NP) *bearded*$_2$ *man*$_1$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{man} \end{array}\right] \end{array}\right] \\ \text{POSSESSOR} & \text{bearded} \\ \text{INDICATOR} & - \end{array}\right]$$

- Possession: 2 has 1.

  (IC) *The nation has*$_2$ *a big debt*$_1$.

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{have} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Stative} \\ \text{FILLER} & \text{nation} \end{array}\right] \end{array}\right] \\ \text{POSSESSION} & \text{debt} \\ \text{INDICATOR} & - \end{array}\right]$$

  (NP) *national*$_2$ *debt*$_1$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{national} \end{array}\right] \end{array}\right] \\ \text{POSSESSION} & \text{debt} \\ \text{INDICATOR} & - \end{array}\right]$$

- Instrument: 1 uses 2.

  (IC) *The system administrator notified*$_1$ *the users via e-mail*$_2$.

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{notify} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{administrator} \end{array}\right] \end{array}\right] \\ \text{INSTRUMENT} & \text{e-mail} \\ \text{INDICATOR} & \text{via} \end{array}\right]$$

  (NP) *e-mail*$_2$ *notification*$_1$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{notification} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & - \end{array}\right] \end{array}\right] \\ \text{INSTRUMENT} & \text{e-mail} \\ \text{INDICATOR} & - \end{array}\right]$$

- OBJECT: 1 is acted upon by 2.

  (IC) $\underline{\textit{They repair}}_2$ $\underline{\textit{engines}}_1$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{repair} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{they} \end{bmatrix} \end{bmatrix} \\ \text{OBJECT} & \text{engine[s]} \\ \text{INDICATOR} & - \end{bmatrix}$$

  (NP) $\underline{\textit{engine}}_1$ $\underline{\textit{repair}}_2$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{repair} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \\ \text{OBJECT} & \text{engine} \\ \text{INDICATOR} & - \end{bmatrix}$$

- OBJECT-PROPERTY: 1 was acted upon by 2.

  (IC) $\underline{\textit{They repaired}}_2$ $\underline{\textit{the engine}}_1$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{repair} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{they} \end{bmatrix} \end{bmatrix} \\ \text{OBJECT-PROPERTY} & \text{[the] engine} \\ \text{INDICATOR} & - \end{bmatrix}$$

  (NP) $\underline{\textit{repaired}}_2$ $\underline{\textit{engine}}_1$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{repair} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & - \end{bmatrix} \end{bmatrix} \\ \text{OBJECT-PROPERTY} & \text{engine} \\ \text{INDICATOR} & - \end{bmatrix}$$

- RECIPIENT: 2 receives the object of 1.

  (IC) $\underline{\textit{We wrote}}_1$ *Smilla a reference letter to* $\underline{\textit{prospective employers}}_2$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{write} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{letter} \end{bmatrix} \end{bmatrix} \\ \text{RECIPIENT} & \text{[prospective] employers} \\ \text{INDICATOR} & \text{to} \end{bmatrix}$$

- PART: 1 is part of 2.

  (NP) $\underline{\textit{the funnel}}_1$ *of* $\underline{\textit{the ship}}_2$

$$\begin{bmatrix} \text{WHOLE} & \text{ship} \\ \text{PART} & \text{funnel} \\ \text{INDICATOR} & \text{of the} \end{bmatrix}$$

- PRODUCT: 1 produces 2.

  (IC) *The factory builds*$_1$ *cars*$_2$.

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & \text{build} \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{factory} \end{bmatrix} \end{bmatrix} \\ \text{PRODUCT} & \text{cars} \\ \text{INDICATOR} & - \end{bmatrix}$$

  (NP) *car*$_2$ *factory*$_1$

$$\begin{bmatrix} \text{OCCURRENCE} & \begin{bmatrix} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \begin{bmatrix} \text{TYPE} & - \\ \text{FILLER} & \text{factory} \end{bmatrix} \end{bmatrix} \\ \text{PRODUCT} & \text{car} \\ \text{INDICATOR} & - \end{bmatrix}$$

- WHOLE: 2 is part of 1.

  (NP) *daisy*$_2$ *chain*$_1$

$$\begin{bmatrix} \text{WHOLE} & \text{chain} \\ \text{PART} & \text{daisy} \\ \text{INDICATOR} & - \end{bmatrix}$$

A.6 Quality

- CONTENT(physical content):1 contains 2.

  (IC) *He filled the bottle*$_1$ *with milk*$_2$.

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{fill} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Container} \\ \text{FILLER} & \text{bottle} \end{array}\right] \end{array}\right] \\ \text{CONTENT} & \text{milk} \\ \text{INDICATOR} & \text{with} \end{array}\right]$$

  (NP) *milk*$_2$ *bottle*$_1$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{bottle} \end{array}\right] \end{array}\right] \\ \text{CONTENT} & \text{milk} \\ \text{INDICATOR} & - \end{array}\right]$$

- TOPIC(abstract content): 1 is concerned with 2.

  (IC) *John produced*$_1$ *a documentary about volcanoes*$_2$.

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{produce} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{documentary} \end{array}\right] \end{array}\right] \\ \text{TOPIC} & \text{volcanoes} \\ \text{INDICATOR} & \text{about} \end{array}\right]$$

  (NP) *volcano*$_2$ *documentary*$_1$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{documentary} \end{array}\right] \end{array}\right] \\ \text{TOPIC} & \text{volcano} \\ \text{INDICATOR} & - \end{array}\right]$$

- CONTAINER: 1 is contained in 2.

  (IC) *He poured milk*$_1$ *into the bottle*$_2$.

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{pour} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Content} \\ \text{FILLER} & \text{milk} \end{array}\right] \end{array}\right] \\ \text{CONTAINER} & \text{bottle} \\ \text{INDICATOR} & \text{into} \end{array}\right]$$

  (NP) *bottle*$_2$ *of milk*$_1$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{milk} \end{array}\right] \end{array}\right] \\ \text{CONTAINER} & \text{bottle} \\ \text{INDICATOR} & \text{of} \end{array}\right]$$

- Manner: 1 occurs in the way indicated by 2.

(CL) *You should write*$_1$ *as I tell you*$_2$.

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{write} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{you} \end{array}\right] \end{array}\right] \\
\text{MANNER} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{tell} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{I} \end{array}\right] \end{array}\right] \\
\text{INDICATOR} & \text{as}
\end{array}\right]$$

(IC) *You write*$_1$ *with style*$_2$.

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{write} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Agent} \\ \text{FILLER} & \text{you} \end{array}\right] \end{array}\right] \\
\text{MANNER} & \text{style} \\
\text{INDICATOR} & \text{with}
\end{array}\right]$$

(NP) *stylish*$_2$ *writing*$_1$

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{writing} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & - \end{array}\right] \end{array}\right] \\
\text{MANNER} & \text{stylish} \\
\text{INDICATOR} & -
\end{array}\right]$$

- Material: 1 is made of 2.

(IC) *We build houses*$_1$ *with bricks*$_2$.

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{build} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{house} \end{array}\right] \end{array}\right] \\
\text{MATERIAL} & \text{brick} \\
\text{INDICATOR} & \text{with}
\end{array}\right]$$

(NP) *brick*$_2$ *houses*$_1$

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{house} \end{array}\right] \end{array}\right] \\
\text{MATERIAL} & \text{brick} \\
\text{INDICATOR} & -
\end{array}\right]$$

- Measure: 2 is a measure of 1

(IC) *The car cost*$_1$ *five hundred dollars*$_2$.

$$\left[\begin{array}{ll}
\text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{cost} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \textsc{Object} \\ \text{FILLER} & \text{car} \end{array}\right] \end{array}\right] \\
\text{MEASURE} & \text{[five hundred] dollar} \\
\text{INDICATOR} & -
\end{array}\right]$$

(NP) *five-hundred dollar*$_2$ *car*$_1$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{car} \end{array}\right] \end{array}\right] \\ \text{MEASURE} & \text{five-hundred dollar} \\ \text{INDICATOR} & - \end{array}\right]$$

- ORDER: 1 is before 2 in a sequence.

  (IC) *He filed the Baker file*$_1$ *before* *the Abel file*$_2$*.*

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & \text{file} \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & \text{OBJECT} \\ \text{FILLER} & \text{[Baker] file} \end{array}\right] \end{array}\right] \\ \text{ORDER} & \text{[Abel] file} \\ \text{INDICATOR} & \text{before} \end{array}\right]$$

  (NP) *A files*$_1$ *before* *B files*$_2$

$$\left[\begin{array}{ll} \text{OCCURRENCE} & \left[\begin{array}{ll} \text{VERB/STATE} & - \\ \text{OCCURRENCE PART} & \left[\begin{array}{ll} \text{TYPE} & - \\ \text{FILLER} & \text{[A] file} \end{array}\right] \end{array}\right] \\ \text{ORDER} & \text{[B] file} \\ \text{INDICATOR} & \text{before} \end{array}\right]$$

- EQUATIVE: 1 is also 2.

  (NP) *composer*$_2$-*arranger*$_1$

$$\left[\begin{array}{ll} \text{ENTITY1} & \text{composer} \\ \text{ENTITY2} & \text{arranger} \\ \text{INDICATOR} & - \end{array}\right]$$

- TYPE: 1 is a type of 2.

  (NP) *oak*$_1$ *tree*$_2$

$$\left[\begin{array}{ll} \text{ENTITY} & \text{tree} \\ \text{TYPE} & \text{oak} \\ \text{INDICATOR} & - \end{array}\right]$$

Appendix B

Machine Learning Results

B.1 C5.0 Accuracy Results

Relation	Baseline	Nr. of	Mean	Standard	Standard
	error	folds	error	deviation	error
Agent	9.52%	5	6.1%	2.0	0.9
Beneficiary	1.43%	4	17.7%	22.6	11.3
Container	0.39%	3	10.8%	9.6	5.5
Content	2.21%	5	23.1%	16.0	7.2
Cause	2.47%	5	32.9%	16.2	7.2
Detraction	0.52%	4	36.3%	12.5	6.2
Direction	1.04%	4	6.2%	8.0	4.0
Effect	4.82%	5	19.0%	9.9	4.4
Equative	2.21%	5	27.7%	16.2	7.3
Frequency	2.21%	5	11.9%	14.8	6.6
Instrument	5.73%	5	42.2%	15.0	6.7
LocationAt	3.12%	5	36.1%	23.3	10.4
LocationFrom	3.65%	5	56.8%	24.3	10.8
Located	0.91%	4	50.2%	39.9	19.9
Material	5.73%	5	19.4%	7.2	3.2
Measure	4.04%	5	16.1%	14.8	6.6
Object-Property	1.95%	5	0.0%	0.0	0.0
Object	5.86%	5	16.9%	18.4	8.2
Part	1.95%	5	13.5%	19.4	8.7
Possessor	5.60%	5	9.1%	5.2	2.3
Product	2.60%	5	16.6%	9.1	4.1
Property	6.78%	5	21.4%	17.1	7.6
Purpose	5.74%	5	50.1%	11.8	5.3
Source	2.74%	5	29.1%	19.1	8.5
Stative	1.43%	4	4.0%	1.4	0.7
TimeAt	4.04%	5	26.1%	11.0	4.9
Topic	7.04%	5	17.6%	12.4	5.5
TimeThrough	0.78%	3	33.8%	36.2	20.9
Type	2.08%	4	8.1%	7.0	3.5
Whole	1.30%	5	9.7%	17.1	7.6

Table B.1: Complete accuracy results obtained in learning rules with C5.0

B.2 Sample Rules Obtained with RIPPER and FOIL

We present in the following sections of this appendix the rules obtained with RIPPER and FOIL, when learning the assignment for the following relations: Agent, Cause, Effect, Frequency, Instrument, LocationAt, Measure, Object, Object-Property to modifier-noun pairs.

In all these experiments we use information about the source of the words (deverbal/gerund/true noun, denominal/deverbal/true adjective or adverb). We also use misclassification costs of 1:3 positive to negative, in order to force the production of more accurate rules. This constraint leads to the production of some rules that do not capture any generalization, but key in on specific words.

B.2.1 Agent

- Rules produced by RIPPER, using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
ag :- head_source=dvn,
      hypernyms_depth_3_mod=person_individual_..._human_soul (50/4).
ag :- hypernyms_depth_2_mod=social_group, head_source=dvn,
      modifier_pos=n (8/0).
ag :- modifier_pos_change=a_n, modifier=national, head_source=dvn (3/1).
ag :- modifier=clerical (2/0).
ag :- modifier=royal (2/0).
ag :- hypernyms_depth_5_mod=atmospheric_electricity (2/0).
ag :- hypernyms_depth_4_head=organic_process_biological_process (2/3).
ag :- modifier=ship (1/0).
ag :- modifier=band (1/0).
ag :- modifier=factory (1/0).
default negative (613/1).
=============================== summary ===============================
Train error rate:  1.30% +/- 0.43% (694 datapoints)    <<
Hypothesis size:   10 rules, 25 conditions
```

- Rules produced by RIPPER, using *Roget's Thesaurus*.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
ag :- head_source=dvn, mod_class=communication_of_ideas (8/6).
ag :- head_source=dvn, modifier_pos_change=a_n,
      head_class=social_volition (7/1).
ag :- head_source=dvn, mod_headword=animality_animal (2/0).
ag :- head_source=dvn, mod_class=individual_volition,
      head_class=individual_volition (5/1).
ag :- head_source=dvn, modifier=national (3/1).
ag :- head_section=results_of_reasoning (2/2).
ag :- modifier=lightning (1/1).
ag :- modifier=band (1/0).
ag :- modifier=royal (1/0).
ag :- modifier=ship (1/0).
ag :- modifier=government (1/1).
ag :- modifier=judicial (1/1).
default negative (471/2).
=============================== summary ===============================
Train error rate:  3.08% +/- 0.76% (520 datapoints)    <<
Hypothesis size:   12 rules, 31 conditions
```

- Rules produced by FOIL, using *WordNet*.

```
is_ag(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        Wordperson_individual_someone_somebody_mortal_human_soul,
        Q,R,S,T,Worddvn,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :- E<>N
is_ag(A,POSn,C,D,E,F,G,H,I,J,K,L,M,N,O,
        P,Wordsocial_group,R,S,T,Worddvn,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_ag(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
        U,V,W,X,Y,Z,AA,Worddecision_making_deciding,AC,AD,AE,AF) :-
is_ag(A,B,C,D,E,F,G,H,I,J,K,L,A,N,O,
        P,Q,R,S,T,U,V,W,X,Y,Z,AA,S,AC,AD,AE,AF) :- A<>N, S<>AC
is_ag(Wordclerical,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_ag(Wordroyal,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_ag(Wordnational,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,Q,AF) :-
is_ag(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,Wordgroup_action,AF) :- W<>AC
is_ag(Wordband,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_ag(Wordship,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

B.2.2 Cause

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
cs :- hypernyms_depth_3_mod=physiological_state (9/2).
cs :- hypernyms_depth_2_mod=happening_occurrence_natural_event (3/3).
cs :- modifier=infectious (1/0).
cs :- modifier=suspense (1/0).
cs :- modifier=growth (1/1).
cs :- modifier=mortal (1/0).
cs :- modifier=tear (1/0).
cs :- modifier=storm (1/0).
default negative (723/1).
============================== summary ===============================
Train error rate:  0.94% +/- 0.35% (748 datapoints)    <<
Hypothesis size:   8 rules, 16 conditions
```

- Rules produced by RIPPER using *Roget's Thesaurus*.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
cs :- mod_headword=ill_health_disease (4/0).
cs :- mod_section=personal_emotion, modifier=tear (1/0).
cs :- modifier=sob (1/0).
cs :- mod_headword=impulse (2/0).
cs :- modifier=mortal (1/0).
cs :- modifier=fertility (1/0).
cs :- modifier=storm (1/0).
cs :- modifier=traumatic (1/1).
cs :- modifier=suspense (1/0).
cs :- modifier=death (1/0).
cs :- modifier=infectious (1/0).
default negative (524/0).
============================== summary ===============================
Train error rate:  0.19% +/- 0.19% (540 datapoints)    <<
Hypothesis size:   11 rules, 23 conditions
```

- Rules produced by FOIL using *WordNet*1.6.

```
is_cs(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        Wordphysiological_state,Q,R,S,T,U,V,W,X,Y,Z,
        AA,AB,AC,AD,AE,Wordentity_something) :-
is_cs(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
        P,Wordhappening_occurrence_natural_event,R,S,T,U,V,W,X,Y,Z,
    AA,AB,AC,AD,AE,AF) :- S<>W
is_cs(Wordstorm,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
    P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_cs(Wordsob,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
    P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_cs(Wordsuspense,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
    P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_cs(Wordtear,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
    P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_cs(Wordmortal,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
    P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_cs(Wordinfectious,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
    P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_cs(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
    P,Q,R,Wordevent,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_cs(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
    P,Q,R,Wordhormone,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

B.2.3 Effect

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
eff :- hypernyms_depth_2_head=condition_status,
       hypernyms_depth_4_head=ill_health_..._health_problem (7/1).
eff :- hypernyms_depth_2_head=happening_occurrence_natural_event,
       head_source=dvn (6/1).
eff :- hypernyms_depth_2_head=condition_status, modifier_source=a (4/1).
eff :- hypernyms_depth_6_head=opening_gap (2/0).
eff :- hypernyms_depth_2_head=attribute, modifier=coffee (2/0).
eff :- hypernyms_depth_5_head=symptom (2/1).
eff :- modifier=occupational (1/0).
eff :- modifier=job (1/0).
eff :- head=obstruction (1/0).
eff :- modifier=onion (1/0).
eff :- hypernyms_depth_2_head=feeling, modifier_pos=n (2/2).
eff :- modifier=anode (1/0).
eff :- modifier=drug (1/0).
eff :- head=mark (1/0).
eff :- head=shock (1/0).
default negative (687/4).
=============================== summary ================================
Train error rate:  1.37% +/- 0.43% (730 datapoints)    <<
Hypothesis size:   15 rules, 35 conditions
```

- Rules produced by RIPPER using *Roget's Thesaurus*.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
eff :- mod_class=emotion_religion_and_morality, mod_section=general (4/0).
eff :- mod_changed_word=heat (3/1).
eff :- head=disease (2/0).
eff :- head_class=emotion_religion_and_morality,
       head_headword=suffering (2/0).
eff :- mod_paragraph_keyword=job (2/0).
eff :- head=shock (2/0).
eff :- modifier=fatigue (1/0).
eff :- modifier=anode (1/0).
eff :- modifier=bow (1/0).
eff :- head=pressure (1/0).
eff :- modifier=technical (1/0).
eff :- modifier=planetary (1/1).
eff :- modifier=traumatic (1/1).
eff :- head=mark (1/0).
eff :- head=burn (1/0).
default negative (490/7).
=============================== summary ===============================
Train error rate:  1.91% +/- 0.60% (524 datapoints)    <<
Hypothesis size:   15 rules, 32 conditions
```

- Rules produced by FOIL using *WordNet*1.6.

```
is_eff(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
         P,Q,R,S,T,U,V,W,X,Y,Z,
         AA,AB,AC,AD,Wordcondition_status,AF) :- W<>AB
is_eff(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
         P,Q,R,S,T,Worddvn,V,W,X,Y,Z,AA,AB,AC,AD,
         Wordhappening_occurrence_natural_event,AF) :- A<>Wordforward
is_eff(Wordcoffee,B,C,D,E,F,G,H,I,J,K,L,M,N,O,
         P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,Wordopening_gap,AB,AC,AD,AE,AF) :-
is_eff(Wordexam,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Worddrug,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Wordheat,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Wordlaugh,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Wordonion,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Wordoccupational,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Wordfatigue,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Wordemotional,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Wordanode,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(Wordtechnical,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_eff(A,B,C,D,Wordair,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

B.2.4 Frequency

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
freq :- modifier_pos=r (7/0).
freq :- hypernyms_depth_1_mod=regular_irregular (4/0).
freq :- hypernyms_depth_1_mod=cyclic_noncyclic_cyclical (3/0).
freq :- modifier=yearly (1/0).
freq :- modifier=occasional (1/0).
freq :- modifier=periodic (1/0).
default negative (733/0).
=============================== summary ================================
Train error rate:  0.00% +/- 0.00% (750 datapoints)    <<
Hypothesis size:   6 rules, 12 conditions
```

- Rules produced by RIPPER using *Roget's Thesaurus*.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
freq :- mod_headword=periodicity_regularity_of_recurrence (5/0).
freq :- modifier_pos_change=r_n (4/0).
freq :- modifier=daily (1/0).
freq :- modifier=yearly (1/0).
freq :- modifier=occasional (1/0).
default negative (531/0).
=============================== summary ================================
Train error rate:  0.00% +/- 0.00% (543 datapoints)    <<
Hypothesis size:   5 rules, 10 conditions
```

- Rules produced by FOIL using *WordNet*1.6..

```
is_freq(A,POSr,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
          R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_freq(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
          Wordregular_irregular,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_freq(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
          Wordcyclic_noncyclic_cyclical,S,T,U,V,W,X,Y,Z,
          AA,AB,AC,AD,AE,AF) :-
is_freq(Wordperiodic,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
          R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_freq(Wordoccasional,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
          R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_freq(Wordyearly,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
          R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

B.2.5 Instrument

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
inst :- hypernyms_depth_4_head=instrumentality_instrumentation,
        hypernyms_depth_5_head=device (12/12).
inst :- hypernyms_depth_3_head=artifact_artefact,
        hypernyms_depth_3_mod=physical_phenomenon (3/1).
inst :- hypernyms_depth_2_head=activity,
        hypernyms_depth_4_mod=instrumentality_instrumentation (5/0).
inst :- hypernyms_depth_3_mod=content_cognitive_content_mental_object,
        hypernyms_depth_2_head=activity (3/0).
inst :- hypernyms_depth_2_mod=attribute, head_source=dvn (3/1).
inst :- hypernyms_depth_7_head=home_appliance_household_appliance (3/0).
inst :- hypernyms_depth_2_head=relation,
        hypernyms_depth_4_mod=instrumentality_instrumentation (2/0).
inst :- hypernyms_depth_3_head=control_controlling (2/1).
inst :- hypernyms_depth_3_head=artifact_artefact, head_source=dvn (3/5).
inst :- modifier=affixal (1/0).
inst :- modifier=psychological (1/0).
inst :- modifier=radio (1/0).
inst :- modifier=vacuum (1/0).
inst :- modifier=shock (1/0).
inst :- modifier=nuclear (1/0).
default negative (659/2).
=============================== summary ===============================
Train error rate:  3.04% +/- 0.64% (723 datapoints)    <<
Hypothesis size:   15 rules, 37 conditions
```

- Rules produced by RIPPER using *Roget's Thesaurus*.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
inst :- head_source=dvn, modifier_pos=n,
        mod_section=results_of_reasoning (2/0).
inst :- head_class=individual_volition,
        head_headword=tool, mod_class=matter (3/1).
inst :- mod_headword=power, head_class=individual_volition (2/0).
inst :- head_paragraph_keyword=furnace (2/0).
inst :- head_source=dvn, mod_section=quantity (2/0).
inst :- head_paragraph_keyword=music (3/2).
inst :- head_paragraph_keyword=message (2/0).
inst :- mod_changed_word=electricity (2/4).
inst :- mod_changed_word=voice (1/1).
inst :- modifier=psychological (1/0).
inst :- modifier=machine (1/0).
inst :- modifier=starvation (1/0).
inst :- modifier=wind (1/0).
inst :- modifier=electron (1/0).
inst :- modifier=vacuum (1/0).
inst :- modifier=needle (1/0).
inst :- modifier=smoke (1/0).
inst :- modifier=finger (1/0).
inst :- head=generator (1/0).
inst :- modifier=steam (1/0).
inst :- head=transport (1/0).
inst :- head=wheel (1/0).
inst :- modifier=refrigeration (1/0).
default negative (481/0).
============================== summary ===============================
Train error rate:  1.53% +/- 0.54% (522 datapoints)    <<
Hypothesis size:   23 rules, 52 conditions
```

- Rules produced by FOIL using *WordNet*1.6.

```
is_inst(A,B,C,D,E,F,G,H,I,J,K,L,E,N,O,P,Q,R,S,T,U,
          V,S,X,Y,Z,AA,Worddevice,AC,AD,AE,AF) :- A<>Wordmail
is_inst(A,B,C,D,E,F,G,H,I,J,K,L,M,N,
          Wordinstrumentality_instrumentation,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,Wordactivity,AF) :-
is_inst(Wordfaith,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,
          Wordhome_appliance_household_appliance,AA,AB,AC,AD,AE,AF) :-
is_inst(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,AA,N,
          Wordinstrumentality_instrumentation,AD,AE,AF) :-
is_inst(A,POSn,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,
          Wordinstrument,AB,AC,AD,AE,AF) :-
is_inst(Wordchemical,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordpressure,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordvoice,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordnuclear,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordsmoke,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(A,B,C,D,E,F,G,H,I,J,K,Wordtelecommunication,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :- O<>AC
is_inst(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,AA,AB,AC,
          Wordcontrol_controlling,AE,AF) :- E<>N
is_inst(Wordmachine,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordpsychological,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordshock,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordwind,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordaffixal,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(Wordvacuum,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(A,B,C,D,Wordvocal_instrumental,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,Wordgenerator,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_inst(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,Wordblanket,T,U,
          V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

B.2.6 Location At

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
lat :- hypernyms_depth_3_mod=location, head_source=n,
       hypernyms_depth_4_mod=region (8/8).
lat :- hypernyms_depth_1_mod=high_low (3/0).
lat :- hypernyms_depth_3_mod=body_of_water_water (2/1).
lat :- modifier=internal (1/0).
lat :- modifier=margin (1/0).
lat :- modifier=office (1/0).
lat :- modifier=chest (1/0).
lat :- modifier=west (1/0).
lat :- modifier=nearby (1/0).
lat :- head=lodge (1/0).
lat :- modifier=surface (1/0).
lat :- modifier=terrestrial (1/0).
lat :- modifier=aquatic (1/1).
default negative (709/1).
=============================== summary ===============================
Train error rate:  1.48% +/- 0.44% (743 datapoints)    <<
Hypothesis size:   13 rules, 28 conditions
```

- Rules produced by RIPPER using *Roget's Thesaurus*.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
lat :- head_class=matter, modifier=high (2/0).
lat :- head_section=organic_matter, mod_section=inorganic_matter (4/0).
lat :- mod_section=space_in_general, modifier=neighbourhood (1/0).
lat :- mod_class=space, modifier=chest (1/0).
lat :- mod_changed_word=city (2/3).
lat :- mod_section=dimensions, head_section=space_in_general (2/2).
lat :- modifier=desert (2/0).
lat :- mod_section=dimensions, modifier=west (1/0).
lat :- modifier=office (1/0).
lat :- modifier=upper (1/0).
lat :- modifier=ocean (1/0).
lat :- mod_section=dimensions, modifier=margin (1/0).
lat :- mod_section=dimensions, modifier=surface (1/0).
lat :- modifier=internal (1/0).
lat :- modifier=lab (1/0).
default negative (506/0).
=============================== summary ===============================
Train error rate:  0.94% +/- 0.42% (533 datapoints)    <<
Hypothesis size:   15 rules, 38 conditions
```

- Rules produced by FOIL using *WordNet*1.6.

```
is_lat(A,B,C,D,E,F,G,H,I,J,K,L,M,N,Wordregion,P,Q,R,S,T,
         Wordn,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :- S<>Y, A<>Wordnational
is_lat(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,Wordhigh_low,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,Wordlife,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(A,B,C,D,E,F,G,H,I,J,K,L,M,N,A,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,Wordlocation,AE,AF) :-
is_lat(Wordwest,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(Wordchest,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(Wordlab,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(Wordinternal,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(Wordmargin,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(Wordsurface,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(Wordoffice,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(A,B,C,D,Wordaquatic,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_lat(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,Wordlodge,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

B.2.7 Measure

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
meas :- hypernyms_depth_1_mod=small_large_little_big (12/1).
meas :- modifier=heavy (7/0).
meas :- modifier=thin (2/0).
meas :- modifier=expensive (2/0).
meas :- hypernyms_depth_14_mod=slender_slight_slim (2/0).
meas :- hypernyms_depth_14_mod=short_long (2/0).
meas :- modifier=moderate (1/0).
meas :- modifier=strong (1/0).
meas :- modifier=difficult (1/0).
meas :- modifier=saturation (1/0).
default negative (704/0).
=============================== summary ===============================
Train error rate:  0.14% +/- 0.14% (736 datapoints)    <<
Hypothesis size:   10 rules, 20 conditions
```

- Rules produced by RIPPER using *Roget's Thesaurus.*

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
meas :- modifier_source=a, modifier_pos_change=no,
        mod_section=causation (9/2).
meas :- mod_section=dimensions, mod_headword=size (6/1).
meas :- modifier=tiny (5/0).
meas :- mod_paragraph_keyword=narrow (2/0).
meas :- modifier=thin (2/0).
meas :- modifier=difficult (1/0).
meas :- modifier=saturation (1/0).
meas :- modifier=expensive (1/0).
meas :- modifier=short (1/0).
meas :- modifier=small (1/0).
meas :- modifier=long (1/0).
default negative (492/0).
=============================== summary ===============================
Train error rate:  0.57% +/- 0.33% (525 datapoints)    <<
Hypothesis size:   11 rules, 25 conditions
```

- Rules produced by FOIL using *WordNet.*

```
is_meas(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,
          Wordsmall_large_little_big,S,T,U,V,W,X,Y,Z,
          AA,AB,AC,AD,AE,AF) :- A<>Wordgreat
is_meas(Wordheavy,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_meas(Wordthin,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_meas(Wordexpensive,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_meas(A,B,C,D,Wordslender_slight_slim,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_meas(A,B,C,D,Wordshort_long,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_meas(Wordmoderate,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_meas(Wordstrong,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_meas(Wordsaturation,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_meas(Worddifficult,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
          U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

B.2.8 Object

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
obj :- hypernyms_depth_1_head=act_human_action_human_activity,
       hypernyms_depth_2_mod=object_physical_object,
       hypernyms_depth_2_head=speech_act (4/0).
obj :- head_source=dvn, hypernyms_depth_3_head=change (9/5).
obj :- head_source=dvner, modifier_pos=n,
       hypernyms_depth_4_mod=communication (3/0).
obj :- hypernyms_depth_2_head=activity, modifier=heart (2/0).
obj :- hypernyms_depth_2_head=activity, modifier=urban (2/0).
obj :- head_source=dvner, modifier=census (2/0).
obj :- head_source=dvner, modifier=blood (2/0).
obj :- hypernyms_depth_4_head=creator (4/0).
obj :- head_source=dvn, hypernyms_depth_3_mod=substance_matter (3/0).
obj :- head_source=dvner,
       hypernyms_depth_14_mod=country_state_land_nation (2/0).
obj :- head=abuse (2/0).
obj :- modifier=horse (2/0).
obj :- modifier=birth (1/1).
obj :- modifier=subject (1/0).
obj :- modifier=acoustic (1/0).
obj :- modifier=cover (1/0).
obj :- modifier=dream (1/0).
obj :- modifier=tuition (1/0).
obj :- head=critique (1/1).
default negative (670/1).
=============================== summary ===============================
Train error rate:  1.11% +/- 0.39% (722 datapoints)    <<
Hypothesis size:   19 rules, 49 conditions
```

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
obj :- head_source=dvn, head_headword=restoration (3/0).
obj :- head_source=dvn, modifier_source=n, mod_section=causation (3/1).
obj :- modifier=horse (1/0).
obj :- modifier=heart (2/0).
obj :- head_source=dvn, mod_section=means_of_communicating_ideas (3/3).
obj :- modifier=jungle (1/0).
obj :- modifier=food (1/0).
default negative (497/13).
```

```
=============================== summary ===============================
Train error rate:  3.22% +/- 0.77% (528 datapoints)    <<
Hypothesis size:   7 rules, 18 conditions
```

- Rules produced by FOIL using *WordNet*1.6.

```
is_obj(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,Wordsubstance_matter,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :- U<>Wordn, A<>Wordwood
is_obj(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,Wordchange,AE,AF) :- E<>M
is_obj(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,B,
         Worddvner,V,W,X,Y,S,AA,AB,AC,AD,AE,AF) :-
is_obj(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,Wordcreator,AD,AE,AF) :-
is_obj(Wordhorse,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordheart,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(A,B,C,D,Wordcountry_state_land_nation,F,G,H,I,J,K,L,M,N,O,P,Q,
         R,S,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,
         Wordabuse,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,
         Wordplanning,T,U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordacoustic,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Worddream,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordsubject,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordjungle,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordland,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordtuition,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordliquor,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordbook,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordcover,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(Wordartifact,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
is_obj(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,Wordcontrol,T,
         U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

B.2.9 Object Property

- Rules produced by RIPPER using *WordNet*1.6.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
obj_prop :- modifier_source=dvapp (15/0).
default negative (737/0).
=============================== summary ===============================
Train error rate:  0.00% +/- 0.00% (752 datapoints)    <<
Hypothesis size:   1 rules, 2 conditions
```

- Rules produced by RIPPER using *Roget's Thesaurus*.

```
option: ratio of cost of FP to cost of FN is 0.461538:1.53846
Final hypothesis is:
obj_prop :- modifier_source=dvapp (13/0).
default negative (529/0).
=============================== summary ===============================
Train error rate:  0.00% +/- 0.00% (542 datapoints)    <<
Hypothesis size:   1 rules, 2 conditions
```

- Rules produced by FOIL using *WordNet*1.6.

```
is_obj_prop(A,B,Worddvapp,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,
              U,V,W,X,Y,Z,AA,AB,AC,AD,AE,AF) :-
```

Appendix C

Parts of speech

Class	Type	Part of speech
DETERMINERS	predeterminers	pre_deter
	central determiners	deter
	postdeterminers: cardinal numerals ordinal numerals and general ordinals closed-class quantifiers	card_num ord_num quant
ADJECTIVES	adjectives	adj
NOUNS	count nouns	countnoun
	mass nouns	massnoun
	proper nouns	propernoun
PRONOUNS	personal	pers_pron
	reflexive	reflex_pron
	possessive	poss_pron
	relative	rel_pron
	interrogative	interro_pron
	indefinite	pron
PREPOSITIONS	prepositions	prep
CONJUNCTIONS	conjunctions	conj
ADVEBRS	adverbs	adv
VERBS	auxiliary	aux/be
	modal	modal_aux
	stative	v/stat
	transitive and intransitive	v/tr_intr
	transitive only	v/tr
	intransitive only	v/intr
ADVERBIAL PARTICLES	adverbial particles	adv_particle
SPECIAL	words which do not easily fit into any of the above classes	special_function_word

Table C.1: Parts of speech used by DIPETT