ABSTRACT

Evidential Analysis of Reported Speech

A dissertation presented to the Faculty of the Graduate School of Arts and Sciences of Brandeis University, Waltham, Massachusetts

by Sabine Bergler

This thesis links a linguistic phenomenon, *reported speech*, within a certain context, *newspaper articles*, to a pragmatic interpretation, namely that the complement clause contains the important *primary* information and that the role of the matrix clause is to provide evidence for this primary information. This is captured in the *evidential analysis* proposed in this thesis. The evaluative context of the matrix clause projects up from the lexical semantics of the two major components, namely the *reporting verb* and the *source* (the subject of the reporting clause) and combines with the context of the surrounding text.

This thesis builds on current trends in computational linguistics, taking advanatge of large on-line corpora and developing a lexical semantics. It goes beyond the main focus of current research to automatically extract lexical data, suggesting that a theoretically guided, semi-automatic corpus analysis can yield *unexpected*, (i.e. not automatically detectable) results that can influence the theoretical basis of a lexical semantics.

This thesis introduces a lexical semantics for reporting verbs and some frequent source descriptions within the framework of Pustejovsky's Generative Lexicon theory [Pustejovsky, 1991]. The thesis goes, however, beyond lexical semantics, connecting the lexical properties with a text representation scheme, MTR, which is designed to make explicit the semantic (and pragmatic) information implicit in the reported speech

sentence. Moreover, MTR offers a new way to represent texts introducing a device called *profile structure*, that groups reported speech sentences according to the source. Profiles complement traditional text representation devices such as coherence structure of sentences and temporal structure of events.

This thesis thus spans from lexical semantics to text analysis, suggesting how to integrate current methods and novel devices for a concrete task. This thesis attempts to not only provide new insights into the analysis of reported speech and newspaper articles in general, but also to contribute to the design of more robust NL systems that can analyze real data reliably.

Acknowledgements

My sincere thanks for making this thesis possible go to my parents, my friends, and colleagues who have supported me for so many years and have given me the strength and inspiration to finish this daunting task.

I dedicate this thesis to my parents. To my mother, who sparked in me the curiosity about language and the self confidence to pursue this degree. To my father who trusted and supported me to go my own ways.

James Pustejovsky, as friend, teacher, and advisor, has helped and guided me along. His enthusiasm for computational linguistics, his broad, interdisciplinary approach, and his innovative research have shaped and will continue to inspire my work. I hope that our regular discussions of linguistics, politics, literature, and life will also continue.

My committee members Richard Alterman, David Waltz, and Yorick Wilks have helped me to understand the process of my own research and its consequences. Despite their own tight schedules they found the time for insightful comments and helpful hints.

Ray Jackendoff's comments helped me at a crucial moment. Lawrence Bookman and Marie Meteer read an early draft of the thesis. Their comments were most helpful.

Marie Meteer has been a close friend, fellow student, and colleague since I came to the United States. She showed me that it could be done, she had advice when all seemed impossible, and she encouraged me to trust my intuitions. Special thanks to her and to David McDonald for giving me shelter and a sense of family during the last nine months and being warm and supportive even though I was moody, depressed, frantic, and introverted for most of this time.

Marc Feeley was always there to help and answer my questions. He has made the last three years enjoyable despite all the craziness and I hope I can be as much help to him in finishing his degree as he has been for me.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

One of the major challenges today is coping with an overabundance of potentially important information. With newspapers such as the Wall Street Journal available on-line as a large text data base, the analysis of natural language texts for the purpose of information retrieval has found renewed interest. This is shown by the large TIPSTER effort funded by DARPA that attempts to *detect* appropriate articles from a body of newspaper articles when queried and to *extract* the appropriate information into a database.

Natural language processing and computational linguistics have largely dealt with sentences that assert *facts* which can be validated. But simple indicative sentences do not form the body of real data found in text corpora, such as newswire or newspaper texts. Here, we find that factual statements occur most frequently *embedded* within contexts of belief or reporting. Accounting for the representation of these embedded statements is vital for the successful interpretation of the overall text. Newspaper articles in particular embed almost all factual information in the context of *reported speech*. The current failure to analyze complex articles stems in part from a *mismatch* between the concerns of philosophers and computational linguists. Whereas the former have devised mathematically describable models for modality, beliefs, and speech acts of arbitrary complexity, the latter have focused on designing and implementing systems to efficiently yet systematically evaluate and maintain truth and belief systems for much

simpler systems. The premise of this thesis is that natural language understanding systems have to address more complex constructions (even if only partially resolved) but that most models proposed by language philosophers and theoretical linguists are too general and too inference-intensive to fill the gap.

The proposal of this thesis is to set the parameters of the investigation differently, namely to investigate complex language phenomena but to delimit the complexity by controlling the context in which the phenomenon is studied. The phenomenon under investigation is *reported speech*, which occurs very frequently in newspaper articles but has not found a principled computational treatment to date. The context in which reported speech will be studied is that of *newspaper articles*, the context in which it plays an important functional role which is quite different from the role of reported speech in spoken everyday discourse or literary fiction.

This thesis shows that the function of reported speech in newspaper articles is one of providing evidence for a proposition and therefore requires an *evidential analysis*. The evidential analysis, in short, acknowledges the embedded clause in reported speech as containing *primary information* that advances the point of the (news) story; the matrix clause of reported speech in newspaper articles has to be seen as the *evaluative environment* for that primary information. The evaluation criterion is one of *reliability*: how reliable is the primary information? This evaluation criterion is shown to be a conventionalized feature of *newspaper style* and it is the lexical semantics of the two functional parts of the matrix clause, namely the *source* and the *reporting verb*, that provide the evidence or the *evaluative environment*.

This evidential analysis is the central point of this thesis. The analysis is very simple, yet to my knowledge has not been recognized in the literature so far.

To evaluate whether the evidential analysis proposed here is (a) feasible and (b) useful this thesis investigates a range of different but related phenomena.

Feasibility

To show that the evidential analysis is feasible means to identify the knowledge required to perform such an analysis and to formalize how this knowledge can be derived. The

approach taken here is to firmly ground the evidential analysis in a rich lexical semantics, Pustejovsky's Generative Lexicon theory [Pustejovsky, 1991]. Lexical entries for a substantial set of reporting verbs are derived partly through corpus analysis, partly through the intensive study of dictionaries. The impact of the lexical semantics of the semantic field of reporting verbs in combination with the lexical semantics of the source NPs is to guide and constrain the reasoning process necessary to evaluate reported speech. This process is demonstrated on different examples.

The representation of the semantic field of reporting verbs within the framework of the Generative Lexicon leads to two new meta-lexical devices to structure the semantic field: a *style sheet* summarizes the essential pragmatic and semantic aspects of reported speech in newspaper style that are linked to the individaul lexical entries and their Lexical Conceptual Paradigms through the *semantic class* construct. Style sheets, in particular, allow us to define a clean interface between lexical semantics on one side and conventionalized usage, basic pragmatics (speech acts), and other style dependent factors on the other. Such a clean separation is preferable to implicit information because the style sheet can be tailored exactly to the corpus at hand without any need to adjust the lexical semantics proper.

Usefulness

The usefulness of the evidential analysis is exemplified by way of introducing a new text representation scheme, Minimal Text Representation (MTR). MTR achieves a representation of the linguistic characteristics of a text, focusing on making explicit what the underlying assumptions are. This is the case for reported speech: the evidential analysis makes explicit what the function of the reported speech is.

MTR does not apply any world knowledge or belief base or even a commonsense inferencer and is thus not a representation of the meaning of a text. MTR constitutes an intermediate representation of the text's linguistic properties and can be described as the result of a linguistic preprocessing of the text. The advantage of such a preprocessed version of the text is its *reusability*: the same intermediate representation can be used by inference mechanisms with different belief bases and for different tasks.

The most important representational device introduced by MTR concerns the the-

matic or argumentative structure of a newspaper article. The role of reported speech, namely to provide *evidence* for a statement by citing its source, is frequently reflected in the structure of newspaper articles by citing several statements by several (possibly independent) sources to support the same *point of view*. The individual statements are to be interpreted within the context of this *supporting group* and complement each other. Making *supporting groups* explicit in the representation and contrasting supporting groups that support opposite points of view, so called *opposing supporting groups*, yields an outline of the argumentative structure of the article that escapes coherence structure representations [Hobbs, 1982, Polanyi, 1987, Mann and Thompson, 1986]. Again, as it turns out, it is the lexical semantics of the two functional parts of the matrix clause of the reported speech, i.e. the source and the reporting verb, that plays the crucial role in determining the supporting group structures.

Detailed examples from complex Wall Street Journal articles illustrate the usefulness of this representation scheme built on the evidential analysis of reported speech. Supporting group structure turns out to be an important scoping mechanism that allows individual statements to be evaluated within a local context. This notion of a local context together with a lazy evaluation scheme for the evaluation of the reported speech can yield a fair assessment of the role of a statement even under partial analysis.

1.2 Importance of Reported Speech

Why is reported speech of special importance? Reported speech is a phenomenon that occurs in all styles, in spoken and written discourse, in newspaper articles, fiction, and academic writing. It is especially frequent, however, in newspaper articles. Consider for instance the fragment of the most frequent words in a corpus of Wall Street Journal articles presented in Figure 1.1[1].

"Said" is the word form that holds sixteenth place in this list. Considering that that is only one form of a single reporting verb, we can extrapolate that reported speech is indeed a frequent phenomenon. Many newspaper articles consist of up to 90% reported speech sentences. If we ignore this construction we will not be able to analyze texts

[1]The complete list of the 5000 most frequent words was compiled by Ken Church and has been distributed by the electronic newsgroup "langage naturel".

344284	.	143725	to	52474	that	28931	by
313711	,	122157	a	45015	The	28457	with
273626	.ES	105278	and	42971	is	28084	it
273626	.S	98110	in	36419	said	27891	Mr
265693	the	94125	”	33586	on	27792	at
149693	of	52818	for	31516	%		

Figure 1.1: Most frequent forms in the Wall Street Journal corpus.

from news sources.

Literature in linguistics and language philosophy has focused on one particular aspect of reported speech, namely the fact that it creates an intensional context for the complement; reporting verbs share this property with verbs of cognition (*to think, to believe*, cf. [Quine, 1960]). The problem with intensional contexts is the fact that the content of the complement clause does not obey normal truth functional treatment[2] and therefore elaborate models have been developed to account, for instance, for belief systems.

I claim here that reported speech as it occurs in newspaper articles should not be treated like other verbs of cognition. In newspapers reported speech has a well described function and conveys in its complement important *primary* information, which carries the news story forward. Although in this function reported speech still plays the role of an attitude towards a proposition, it cannot be reconciled with the notion of intensional context. For a detailed discussion see Chapter 3.

The evidential analysis incorporates the function of reported speech to provide evidence for the primary information and thus constrains the inferences necessary to analyze newspaper texts. This is an important gain for the (partial) analysis of large amounts of “real” text.

[2]For more detail see Section 2.2.1.

1.3 Importance of Newspaper Articles

Reported speech in general is a complex problem. The evidential analysis proposed here holds only for one use of reported speech, namely where reported speech is used to attribute a statement to its source in order to indicate its reliability. Is it justified to limit the analysis of a linguistic problem to a particular context?

This thesis claims that it is indeed a successful way to reduce complexity without sacrificing accuracy to study linguistic devices in a narrowly defined context on a large corpus. Delimiting the problem to a specific domain (i.e. newspaper articles) does not necessarily mean simplifying its solution. On the contrary, the solution for reported speech sentences in newspaper articles presented here spans from the lexical semantics of the reporting verb and the source descriptors all the way up to the argumentative structure of the overall text. Only when we consider how all the different steps in text analysis interact and subsequently add to a solution can we be sure to have a feasible model that can be realized in the context of a larger language understanding system. The analysis presented here is thus comprehensive in a different way: instead of studying one phenomenon in all possible contexts on, say, the syntactic level it studies one phenomenon in one context from the lexical semantics up to the text representation level.

But are newspaper articles a useful domain? The answer is yes. The growing availability of electronic news sources makes information retrieval a timely and time consuming task. Current natural language understanding systems are too brittle to handle the incoming texts, while statistical information retrieval methods are too crude and too error prone to satisfy the need for reliable classification of data. Research in both fields is converging on hybrid methods of partial analysis. The traditional way of building a system that successfully analyzes texts of a certain subdomain of topics breaks in the presence of unrelated texts. It is therefore time to adapt the sublanguage approach and develop systems that implement knowledge of certain styles as their "subdomain". Newspaper style is a particularly conventionalized style that can be studied on large on-line corpora, some even available in a tagged version. Knowing style characteristics in turn facilitates the partial analysis of texts when words or constructions are unknown by constraining the possible space of ambiguity.

1.4 The Problem

Truth-functional semantics describes the meaning of a sentence as a mapping into a truth value [Dowty *et al.*, 1981]. Disregarding the question whether that is a satisfactory account of sentence meaning in general it is not sufficient for reported speech as it occurs in newspaper articles. Consider the folowing sentences from the first two paragraphs of a New York Times Article:

> The New York Times, December 8, 1990
>
> The Ford Motor Company and Volkswagen A. G. of Germany are nearing agreement on a joint venture for European production of mini-vans.
>
> ...
>
> The British newspaper The Financial Times said the plant was expected to have a production capacity of 150,000 to 200,000 vehicles a year and could involve an investment of $2.5 billion to $3 billion. Ford and Volkswagen would not comment on those figures.

The reported speech in this text fragment demonstrates why truth is not an adequate notion for the evaluation of reported speech. The truth of the matrix clause is assumed just as with any assertion, yet this does not do justice to the reported speech as a whole. Truth of the complement is a notion that is not only incompatible with all work on intensional contexts, but also with the function of the reported speech here: if the meaning was simply to assert that the complement is true, it would suffice to assert the complement. The function of the reported speech here is clearly to indicate that the reporter cannot or will not evaluate the truth of the complement, yet that its content is important for the text overall. Reported speech allows the reporter to include information that is uncertain, or not necessarily reliable, without committing to its content. By attributing the complement clause to the source of the information, the reporter introduces an evaluative environment in which the reader can assess the reliability of the information. The assessment of reliability is necessarily subjective: depending on knowledge, beliefs, points of view, and interest different readers will evaluate the same instance of reported speech differently.

The problem of determining the meaning of reported speech is then one of determining the relevance and contribution of the primary information in the complement clause

and to evaluate its reliability. But how can reliability be characterized, given that it is basically a subjective measure?

Studying many instances of reported speech in different newspapers on large corpora we find that the reporter encodes the intended reliability in the lexicalization of the source and the reporting verb (additional information may be added about the circumstances of the original utterance by the source). The reader's task is then constrained by the encoded reliability. To recover these constraints requires a rich lexical semantics and composition procedures from the lexical composition of the individual clauses up to the compositionality of the text. This thesis outlines one way to approach this problem.

1.5 The Solution

Given the many layers of problems just outlined there are also many layers of solutions. This section addresses the solution to the problem of the function and analysis of reported speech as a linguistic construct.

Reported speech shares the characteristic of providing evidence for a statement with other linguistic constructs called *evidentials* [Chafe, 1986]. The evidential construct that is closest to reported speech is *hearsay*. A detailed discussion of evidentials can be found in Chapter 3, Section 3.1. Evidentials as a class encode different modes of uncertainty about propositions. The realization that the function of reported speech in newspaper articles, namely to provide evidence, is not unique is an important step towards a solution that can be extended to those other constructs later on. The evidential analysis proposed here is just such a general solution.

The formalization of the evidential analysis hinges on the realization that the information of the complement clause is usually the more important information, the *primary* information, that pushes the story forward. The information of the matrix clause, in contrast, constitutes the evidence of reliability of the primary information, giving the circumstances under which the original utterance was made. This *circumstantial* information consists minimally of the source and a reporting verb but it can also include time and location of the utterance, manner, intension, purpose, and many more. This additional information is added in form of prepositional phrase attachments

or modifiers to the source or reporting verb.

Requirements for the correct analysis of reported speech in newspaper articles are (a) to bring the primary information to the forefront and (b) to determine exactly what the evaluative context is. (a) is a problem of *representation*, (b) one of the (lexical) semantics of the matrix clause. Let me discuss the latter first.

1.5.1 Lexical Semantics of the Matrix Clause

The matrix clause consists of two functional parts, the *source* description and the *reporting verb*. The lexicalization of these two parts determines the evaluative context for the embedded primary information, determining its reliability.

Reporting verbs here are defined as those utterance and perception verbs, that can be used to convey the language of others. This is a narrow interpretation of reporting verbs. I consider for instance *say, tell, insist, announce, claim, deny* as reporting verbs. Verbs of cognition such as *believe, think, know* are not considered reporting verbs for the purposes of this thesis. Reporting verbs are those verbs that report *the speech of others*.

I will further concentrate on reporting verbs that occur frequently in newspaper articles. Examples from the Wall Street Journal corpus are:

(1) (a) Texas Instruments Japan Ltd., a unit of Texas Instruments Inc., said it opened a plant in South Korea to manufacture control devices.
(b) The U.S., claiming some success in its trade diplomacy, removed South Korea, Taiwan and Saudi Arabia from a list of countries it is closely watching for allegedly failing to honor U.S. patents, copyrights and other intellectual-property rights.
(c) The Internal Revenue Service has threatened criminal sanctions against lawyers who fail to report detailed information about clients who pay them more than $10,000 in cash.
(d) Backe Group Inc. agreed to acquire Atlantic Publications Inc., which has 30 community papers and annual sales of $7 million.

Say is the most unmarked of these reporting verbs and only indicates a source for the statement of the complement clause. *Claim*, however, lexically encodes more room of doubt; in fact Webster's Ninth New Collegiate Dictionary defines

claim 3 a. to assert in the face of possible contradiction: maintain

Thus when a reporter uses *claim* as reporting verb, he or she also indicates room for doubt (due to additional knowledge that may or may not be indicated later on). *Agree to*, on the other hand, is an encoding of a (possibly legal) speech act and therefore lexically encodes more reliability (for the pragmatical reason that the reporter and the newspaper could be sued in case of misrepresentation). This short discussion shows that the lexical semantics of the reporting verb is of great importance for the evaluation of the reported primary information. The source description, likewise, plays an important role. In the Wall Street Journal this role is usually to indicate a possible special interest or expertise, as in

(2) (a) Here is the mature Abraham Lincoln uttering a trim three paragraphs — that eventually filled an entire wall of the Lincoln Memorial — at the dedication of a military cemetery in Gettysburg.
(b) The Toronto-based real estate concern said each bond warrant entitles the holder to buy C$1,000 principal amount of debentures at par plus accrued interest to the date of purchase.
(c) Another small Burgundy estate, Coche-Dury, has just offered its 1987 Corton-Charlemagne for $155.

The correct lexical semantics can however not simply be looked up in a dictionary; extensive corpus analysis on a corpus of real data of the appropriate kind (i.e. newspaper articles for the analysis of newspaper articles) has to determine the actual *usage* of a reporting verb in that context. Lexicographers often do not distinguish the reporting verb sense of an utterance verb from other (sometimes more literal) word senses. This thesis will go to the other extreme and only discuss the reporting verb senses for the verbs considered.

1.5.2 Representation

Once the evaluative environment for the complement clause has been determined, it has to be represented in a way that brings the primary information to the forefront and does not loose the information of the evaluative environment. Chapter 3 introduces the following notation:

S'[OC,CC,TC]

where

S' is the interpretation of the (reported) clause,
OC is a variable for the original context,
CC is a variable for the current context, and
TC is a variable for the temporal context.

OC, CC, and TC are *context variables*, which indicate the original context (OC), as encoded in the reporting clause, the current context (CC), that is the position in the text structure or argument structure of the text, and the temporal context (TC), that is the position in the trace.[3]

This representation inverses traditional predicate calculus representations of complex clauses by making the matrix clause an (environmental) variable of the complement clause. This representation is part of a larger text representation scheme at an intermediate level, where information implicit in linguistic devices has been made explicit but no commonsense inferences or points of view have been brought to bear, effectively achieving a Minimal Text Representation (MTR) that can serve as a *resource* for several different *evaluation methods*; a conservative belief attribution system such as Ballim and Wilk's ViewGen [Ballim and Wilks, 1992], for example, could use an MTR text representation as a basis of belief acquisition.

1.6 An Example

To illustrate on a very short example what the evidential analysis looks like, consider again the text from the New York Times:

The New York Times, December 8, 1990

The Ford Motor Company and Volkswagen A. G. of Germany are nearing agreement on a joint venture for European production of mini-vans.

...

[3]The notation is not only applicable to reported speech contexts and other constructs spanning orthogonal contexts, but can also be useful for simple sentences, where the OC could contain the reference for anaphora etc.

> The British newspaper The Financial Times said the plant was expected to have a production capacity of 150,000 to 200,000 vehicles a year and could involve an investment of $2.5 billion to $3 billion. Ford and Volkswagen would not comment on those figures.

The representation of the reported speech sentence would be the following:

$S_2 = C_2(R_2)$, where

C_2 = HYPOTHESIS(AND(C_2', C_2"))
C_2' = PRODUCTION-CAPACITY($PLANT_1$,
RANGE(150000 vehicles, 200000 vehicles))
C_2" = INVESTMENT(Ford and Volkswagen, $PLANT_1$,
RANGE($2.5 billion, $3 billion))
R_2 = (**OC**: (S: "The British newspaper The Financial Times",
U: say,
C: -),
CC: NIL,
TC: PAST-TENSE(C_2) and PAST-TENSE(R_2))

A more elaborate example together with different scenarios that make use of this representation are presented in Chapter 3, Section 3.5.

1.7 Lexical Semantic Constraints on Reliability

As I have pointed out in Section 1.1, the evidential analysis must be feasible. The previous section explained the formal notation of the evidential analysis but it did not address *how* reliability can be computed from the context variables. One point of this thesis is that the process of assessing reliability is subjective and up to the reader. Yet this thesis also claims that the reader's task is highly constrained by the lexical semantics of the circumstantial information. A major part of this thesis is therefore devoted to deriving a lexical semantics suitable to this task.

I am assuming the framework of the Generative Lexicon here. The *Generative Lexicon* is a formalism currently being extended at Brandeis University based on [Pustejovsky, 1991]. The goal is to provide a representation language that can adequately

describe a lexical semantics that not only defines the meaning(s) of a word but also provides a structure that allows for dynamic construction of clusters of related words — antonyms, synonyms, collocational information and more. The Generative Lexicon (GL) is an interface between linguistic (i.e. syntactic and semantic) terminology and commonsense and real world concepts, enabling independent yet integrated processing of real texts on these different levels. In the philosophy of GL a richer lexical semantics yields a more compositional semantics that can project from the lexical level upwards, even onto the level of *text*.

The richer structure of the Generative Lexicon makes it a useful representation for lexical semantics derived from corpus analysis. Theoretically guided semi-automatic corpus analysis can provide insights into regularities of *usage* that are not usually found in dictionary entries. One example discussed in detail in Chapter 6, Section 6.4.2 is the case of preferences for different kinds of metonymy. Closer to the issue of assessing reliability is the requirement for certain reporting verbs to stand in a context of opposition. These *discourse polarity items* implicitly encode that the primary information is not above challenge, but has in fact already been placed in a context of opposition. *Insist* is such a discourse polarity item. The possibility of doubt expressed by the reporting verb plays an important role in the reliability assessment. Contrast *announce*, a strong, factual verb that lexically encodes a *legitimation* of the source to speak on the topic with *insist*, a discourse polarity item that lexically encodes that the primary information has been challenged before. While this thesis does not assign values of reliability to individual words, the investigations are highly suggestive for a relative ordering of reporting verbs. A comparative study of the structure of the field of reporting verbs in two dictionaries reveals similarites that were used to formulate a set of *semantic dimensions* of this semantic field.

All these investigations into the lexical semantics of individual reporting verbs and their structure as a semantic field finally give rise to a new meta-lexical concept, the concept of a *style sheet*.

The problem that the style sheet addresses is: what is the status of the style dependent aspects of reporting verbs? They are clearly lexical in nature, yet do not belong in a general lexical entry. The solution proposed here is to define two new meta-lexical constructs, called *style sheet* and *semantic class*. The style sheet contains the pragmatic

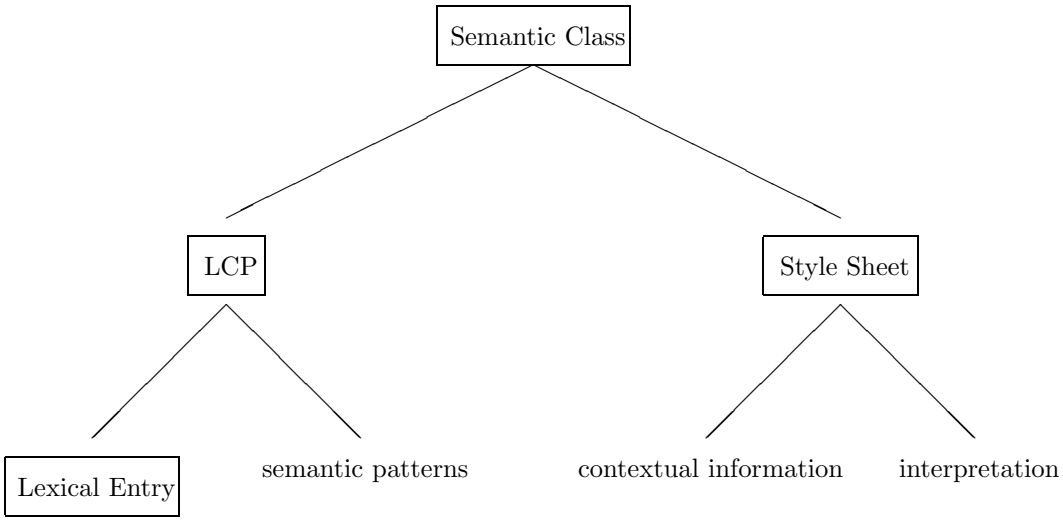


Figure 1.2: Meta-lexical structure.

information associated with reporting verbs in newspaper articles. The semantic class links specific LCPs with the appropriate style sheets. The resulting structure is shown in Figure 1.2.

This layered approach provides a clean interface for the correlation of lexical semantic and pragmatic information. The advantage of the style sheet is that it can be changed or augmented without affecting the basic lexicon.

1.8 Implications for Text Representation

The very principle that the reporter applies in not evaluating the reliability of some statement when using reported speech should be accepted by a text representation. It is therefore important to outline a text representation that can implement an evidential analysis meaningfully, that is a representation that does not force an evaluation. The last two chapters of this thesis introduce a representation scheme that is designed to represent only the linguistically encoded information. This representation scheme, called Minimal Text Representation (MTR), employs three different devices to represent a text. This is fundamentally different from traditional text representations, where one representation combines all aspects of the text. Chapter 3, Section 3.4.1 compares

MTR to Kamp's Discourse Representation Theory (DRT) [Kamp, 1988, Kamp and Reyle, 1991]. There I claim that a representation that disentangles different aspects of the text such as coherence structure and temporal structure will help user programs to efficiently access the representation for their purposes.

The three devices are *coherence structure, trace*, and *profile structure*:

Coherence Structure: Coherence structure records the original structure of the text and the coherence relations between sentences.

Trace: A trace captures the sequence of temporal activities which can be put into a partial order on the narrative time line. A trace is constructed from *trace segments* in a bottom-up fashion using *trace composition* and *trace unification*.

Profile Structure: A profile contains a list of all properties the text asserts or implies about a particular discourse entity. Distinct discourse entities have separate profiles. Profiles can be nested and form a complex profile structure.

Coherence structure is the device that has found most attention in discourse analysis literature (cf. [Polanyi, 1987, Hobbs, 1982, Mann and Thompson, 1986]), capturing the relations that hold between sentences, paragraphs, etc. While coherence structure is a necessary ingredient in text representation, it is not sufficient. An orthogonal structure is the temporal structure of events mentioned in the text. This structure is not easily combined with the text centered coherence structure and is therefore represented separately in MTR in form of partial orderings of the events in form of a trace (cf. [Bergler and Pustejovsky, 1990]).[4]

The third device, profiles, is similarly not a novel concept. Novel is the particular structure that results from the functional analysis of reported speech in newspaper articles, where individuals are grouped according to the ensuing argumentative structure of the text.

[4]In DRT [Kamp, 1988, Kamp and Reyle, 1991] both text structure and temporal information are combined. To separate the two is therefore not a necessity, but an expedient reducing complexity. Because MTR is not intended to yield the final representation of a text, it has to be "parsed" for final interpretation, which is facilitated by separating different information.

It is not the purpose of this thesis to introduce a fully formalized text representation formalism. I will focus here on the representational device that is particularly suitable to represent the function of reported speech in newspaper articles, the *profile structure*.

Let me illustrate these concepts with an example. Consider the following text from the Wall Street Journal:

> **Who's News: Pacific Enterprises Chooses Ukropina As Chief Executive**
> By Jeff Rowe
>
> (S_1) Pacific Enterprises named its president, James R. Ukropina, as chairman and chief executive, succeeding Paul A. Miller and ending a century of family leadership at the utility holding company.
>
> (S_2) Analysts said *(C_1) the naming of Mr. Ukropina represented a conservative move by an unusually conservative utility concern.* (S_3) *(C_2) Unlike some companies, Pacific Enterprises has "made no major errors moving outside their area of expertise,"* said Craig Schwerdt, an analyst with Seidler Amdec Securities Inc. in Los Angeles.
>
> (S_4) *(C_3) "Each of the company's businesses are positioned to do well in the coming year,"* said Paul Milbauer, an analyst with C.J. Lawrence, Morgan Grenfell in New York. (S_5) *(C_4) Most of the company's retail operations are in the fast-growing West, and the gas unit will benefit from tightening environmental regulations,* he said. (S_6) He added that *(C_6) more-stringent pollution controls are expected to increase demand for gas, which is relatively clean-burning.*
>
> Wall Street Journal, 10/5/89

The coherence structure of this text is — in its simplified form — represented in Figure 1.3.

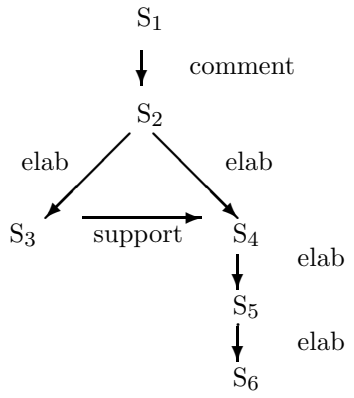


Figure 1.3: Simplified text structure

Sentence S_1 sets the topic of the article. Sentence S_2 presents an evaluation of S_1 (the gist of the evaluation only becomes clear after the third sentence, however). Sentences S_3 – S_6 serve to elaborate (and support) S_2.

The trace is a set of intersecting trace segments representing partially ordered events, illustrated in Figure 1.4.

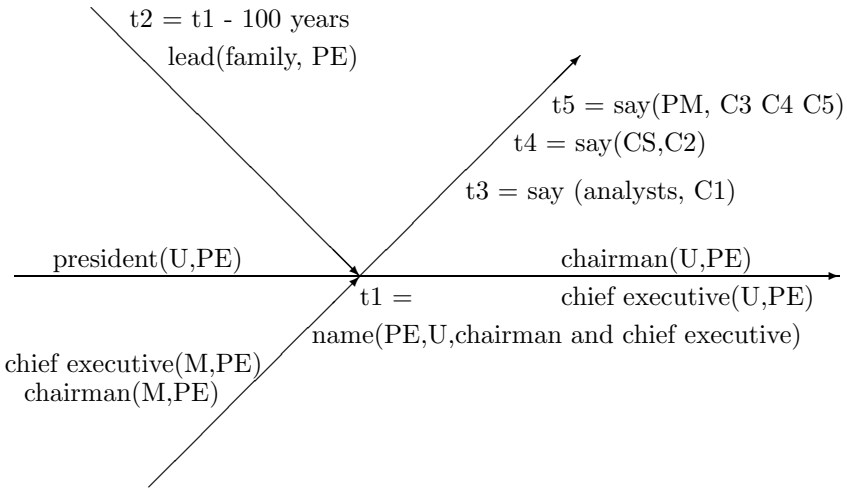


Figure 1.4: Trace

Note that the trace segments have not been unified — the extra inferencing needed to unify the trace segments can be done at any time.

Profiles are a collection of all properties mentioned in a text about an entity, in particular all utterances made by that entity. I will adopt the notation of nested boxes that has been used frequently for similar purposes [Fauconnier, 1985, Kamp, 1988, Ballim and Wilks, 1992].

Utterance Profiles for Ukropina text

Pacific Enterprises
named its president, James R. Ukropina, as chairman and chief executive, succeeding Paul A. Miller and ending a century of family leadership at the utility holding company.

Analysts

said: the naming of Mr. Ukropina represented a conservative move by an unusually conservative utility concern.

Craig Schwerdt, analyst with Seidler Amdec Securities Inc. in Los Angeles

said: Unlike some companies, Pacific Enterprises has "made no major errors moving outside their area of expertise."

Paul Milbauer, analyst with C.J. Lawrence, Morgan Grenfell in New York

said: "Each of the company's businesses are positioned to do well in the coming year."

said: Most of the company's retail operations are in the fast-growing West, and the gas unit will benefit from tightening environmental regulations.

added: that more stringent pollution controls are expected to increase demand for gas, which is relatively clean-burning.

A notion that evolved from that of profile structure is the concept of *supporting group structure*. When an article cites several people in favor of the same issue, possibly contrasting their statements with several other people's who are not in favor of this issue, then we can say that for the sake of the argumentation in the article at hand, these profiles form a group, where the statement of one individual holds in some sense for the whole group. The statements *support* one another and I call the resulting grouping a *supporting group*.

The example text here exhibits such a supporting group, namely the analyst's statements support each other. The representation of supporting groups is a box containing all supporting profiles:

Supporting group of analysts

Analysts

said the naming of Mr. Ukropina represented a conservative move by an unusually conservative utility concern.

Craig Schwerdt, analyst with Seidler Amdec Securities Inc. in Los Angeles

said: Unlike some companies, Pacific Enterprises has "made no major errors moving outside their area of expertise."

Paul Milbauer, analyst with C.J. Lawrence, Morgan Grenfell in New York

said: "Each of the company's businesses are positioned to do well in the coming year."

...

Supporting groups are an important part of the structure of a text. In those newspaper texts where different sources are mentioned it is important to be able to group the arguments being put forth. From the perspective of supporting group analysis there are three different forms of articles, namely

1. Articles that quote from a single source.

2. Articles that quote different sources to support a point.

3. Articles that report two or more different points of view, supporting each with quotes from different sources.

In consequence, there is a notion for supporting groups that support opposing views, so called *opposing supporting groups*. Chapter 9 elaborates these notions in detail and gives a complex example of opposing supporting group structure.

MTR and in particular supporting group structure turns out to be a useful representation even when the utterances are not fully evaluated. The supporting groups provide a *local context*, which allows a user program to evaluate a particular statement in that local context without having to evaluate the whole text.

1.9 Summary

This thesis links a linguistic phenomenon, *reported speech*, within a certain context, *newspaper articles*, to a pragmatic interpretation, namely that reported speech introduces a *context split* into a text. The context of the reported clause is clearly not the surrounding text, but the original discourse situation, or *original context* which is partially represented in the matrix clause. The two levels of information correspond to two distinct contexts, where the matrix clause constitutes a modification of the primary information in the embedded clause with supporting evidence. The circumstantial information of the reporting clause provides in fact an *evaluative context* in which the primary information has to be *evaluated by the reader*. The evaluative context of the reporting clause projects up from the lexical semantics of the two major components, namely the *reporting verb* and the *source* of the information (the subject of the reporting clause) and combines with the context of the surrounding text. Thus, the evaluation of the reporting clause consists of two components, namely a *lexical semantic* and an *inferential* component.

This is a novel approach, departing from traditional computational linguistics. This thesis builds on current trends in computational linguistics, taking advantage of large on-line corpora and developing a lexical semantics. But it goes beyond the main focus of current research in these areas. The main goal in corpus analysis is currently to populate or enhance a computational lexicon with phrasal patterns (cf. [Smadja and McKeown, 1990]) and collocational data (cf. [Hindle, 1990]). I demonstrate, in contrast, that a theoretically guided, semi-automatic corpus analysis can yield *unexpected*, (i.e. not automatically detectable) results that can influence the theoretical basis of a lexical semantics. The example given in Chapter 6 is the case of metonymic behavior of the subject of reporting verbs.

A lexical semantics for reporting verbs is introduced and some frequent source descriptions within the framework of Pustejovsky's Generative Lexicon theory [Pustejovsky, 1991]. Chapters 8 and 9 then go beyond lexical semantics, connecting the lexical properties with a text representation scheme, MTR, which is designed to make explicit the semantic (and pragmatic) information implicit in the reported speech sentence. Moreover, MTR offers a new way to represent texts introducing a device called

profile structure, that groups reported speech sentences according to the source. Profiles complement traditional text representation devices such as coherence structure of sentences and temporal structure of events.

The thesis is divided into nine chapters. Chapter 2.2 introduces the phenomenon of reported speech and surveys briefly some of the linguistic literature where it has been addressed. Chapter 3 introduces literature on *evidentials*, drawing parallels between evidentials in general and reported speech in newspaper articles and working out in detail what constitutes an *evidential analysis* of reported speech.

The following chapters are less theoretical and ground the analysis of the preceeding chapters firmly in data. Chapter 4 analyzes the style of newspaper articles and identifies some constraints on the interpretation of reported speech in this domain that are conventionalized. Chapter 5 introduces the paradigm of lexical semantics which I adopt and presents the methodology applied. Chapter 6 reviews some current techniques of corpus analysis and details on the analysis of *insist* some intricate lexical semantic features that can be extracted from corpora, concerning metonymic behavior and presuppositions. Chapter 7 integrates the points made in the previous chapters in an analysis of the semantic field of reporting verbs, applying the lexical semantics outlined in Chapter 5.

The last two chapters introduce the representation mechanism, *Minimal Text Representation*, that captures the special problems arising from the evidential analysis of reported speech. Chapter 8 introduces the philosophy behind the mechanism and defines its terms (including supporting groups), pointing out its impact on text analysis on several examples taken from on-line corpora. Chapter 9 finally presents a detailed analysis of one text with sufficient complexity to show the workings of the representation mechanism.

Chapter 2

Background

This chapter intruduces some characteristics of reported speech and some literature that has addressed reported speech in different contexts.

2.1 Grammatical Features

Quirk et al remark that the structural relationship between reporting clause and reported clause varies. In some contexts the reported material functions as subordinate clause (Dorothy said, *'My mother is on the phone.'*)[1] or it can be made into a subject complement (What Dorothy said was *'My mother's on the phone.'*), in both cases the complement is obligatory. The reporting clause in both cases appears as an obligatory part because it can be coordinated with a following clause ('The radio is too loud,' *Elizabeth complained*, and she then stalked back to her room.)

In other contexts the reporting clause can be viewed as subordinate, functioning as an adverbial, resembling one type of *comment clause*, that behaves like the matrix clause of a main clause (There were no other applicants, *I believe,* for that job.) Compare [Quirk *et al.*, 1985, p. 1023]:

1. 'Generals,' *they alleged,* 'never retire; they merely fade away.' (reporting clause)

[1]The examples stem from Quirk *et al.*

2. 'Generals,' *it is alleged,* 'never retire; they merely fade away.' (comment clause)

3. 'Generals,' *allegedly,* 'never retire; they merely fade away.' (adverb)

Indirect speech has the interesting property that *deictic* expressions have to be changed appropriately. These include *tense* and *temporal reference* (such as *yesterday*), *pronouns*, *spatial deixis* (such as *here*), and *demonstratives* (such as *this*).

The backshifting behavior for tenses is interesting and complex. Jespersen [Jespersen, 1924] [Jespersen, 1949] presents a very detailed account of backshifting behavior with examples. Backshifting is largely optional, except in present tense, where an unshifted present tense implies that the reported clause is still valid or that the source of the reported clause is famous (*The Bible says ...*). Communications in the recent past (*John tells me ...*) and reports of cognitive processes (*I think ...*) may also occur in present tense in reported speech. All other cases of present tense have to be backshifted. Backshift of the past is optional. The general pattern according to Quirk *et al* is:

DIRECT SPEECH	BACKSHIFTED IN INDIRECT SPEECH
present	∼ past
past	∼ past or past perfective
present perfective	∼ past perfective
past perfective	∼ past perfective

2.2 Reported Speech in the Literature

Reported speech is part of several larger fields of study in linguistics and the philosophy of language. Most importantly, reported speech is related to propositional attitudes, sentences that are not merely factual statements but express the speaker's attitude toward the embedded statement. The prototypical attitude is *believe*, a marker of uncertainty that creates an *intensional context*, which describes a mental state of the speaker but may not correspond to any state in the "real" world.

This section introduces some of the literature that has considered the language philosophical aspects of reported speech in connection with *propositional attitudes, point of view and belief maintenance,* and *speech acts.* The discussion of the literature on these

topics is short and general, serving mainly to introduce the concerns of these fields and how they pertain to my analysis.

2.2.1 Propositional Attitudes

Work on propositional attitudes in linguistics and philosophy has focused on attitudes introduced by verbs of cognition, i.e. *beliefs, desires, expectations, thoughts* (cf. [Quine, 1960]). Reporting verbs are usually subsumed in this category.

Propositional attitude verbs introduce an intensional context for the embedded complement, thus proving problematic for truth-theoretical accounts of semantics, as is the case for the imaginary entity *a unicorn* in (1):

(1) John seeks a unicorn.

Possible world semantics allows to formulate a truth-theoretical account of such sentences by specifying that John's belief state must be such that in addition to all the conditions that are known (and accepted by John) as valid in the current world he also believes that there exist entities that are called unicorns and that they can be sought (contrary to fact). A world corresponding to John's belief state is *possible* but not compatible with the currently *valid* world.

A serious problem with intensional contexts is *opacity*. Opacity prohibits substitutability, that is the referring expression in the embedded clause cannot necessarily be replaced with another referring expression denoting the same underlying entity. Thus (2b) does not necessarily follow from (2a):

(2) (a) John expects the Evening Star to rise in five minutes.
(b) John expects the Morning Star to rise in five minutes.

Opacity is obvious for quotations. Quoted material does not refer in the same way other words and phrases refer and requires therefore special treatment. The distinction is usually glossed as *use vs. mention* of language, where quotations only *mention*, but do not *use* the language enclosed in quotation marks. Thus even though *my wife* in

(3a) refers to *Jocasta,* this fact cannot be exploited when quoting[2] and thus (3a) is not equivalent to (3b).

(3) (a) Oedipus said: "My wife is beautiful."
(b) Oedipus said: "Jocasta is beautiful."

Four different schemes have been proposed in the literature. In the words of Davidson [Davidson, 1979] these are the 'proper-name theory', the 'picture theory', the 'spelling theory', and the 'demonstrative theory' of quotation.

The first three theories refer to different awkward constructions of the quotation as an unstructured, single term ([Quine, 1953]); as a name for the sign or expression inside the quotation marks [Quine, 1940]); and a decomposition of the quotation into a structure of individual words in quotation marks that are concatenated ([Geach, 1972]). Davidson himself favors the 'demonstrative theory', treating the quoted material as semantically irrelevant to the sentence itself (so that truth can be assigned to the sentence) that could at any point in time be replaced by a demonstrative pronoun and a gesture indicating the replaced material ([Davidson, 1979]). The analysis of quotation as demonstrative has recently been elaborated on by [Clark and Gerrig, 1990]. What is common to all these approaches is to consider quotation as a device that is fundamentally different from indirect reported speech, where the quoted material occurs paraphrased and definitely 'used', i.e. semantically relevant in the sentence. I propose that quotation should get the same semantic treatment as does indirect reported speech and will come back to that point.

Another problem with propositional attitudes in general, and indirect speech in particular, is the distinction between *de re* and *de dicto* reference.

(4) (a) Oedipus said that his wife was beautiful.
(b) Oedipus said that his mother was beautiful.

The problem is to determine whether a reference (here *his mother* and *his wife*) in the reported clause is according to what the original source *said* (de dicto) or what the reporter *substituted* for clarification, based on facts he or she knows to be true of the

[2]For a somewhat looser definition of quotation in marked contexts see [Clark and Gerrig, 1990].

entity denoted by the original reference (de re). Thus (4a) is de dicto, (4b) is de re. A complicating factor is that references de dicto do not necessarily have to describe the denoted entity correctly, in conversation (and for communicative purposes) anything that conveys to the hearer *whom* the speaker is referring to is appropriate. I will argue later that this distinction is not important for the analysis of newspaper articles because journalists take care to mark de dicto references clearly.

Davidson's "On Saying That" [Davidson, 1968] also gives a 'demonstrative' solution to the analysis of indirect speech, where he suggests to analyze reported speech as a notational variant of a parataxis in the sense of (5):

(5) (a) Galileo said that the earth moves.
(b) Galileo said that. The earth moves.

That in (5b) functions as a demonstrative pronoun (somewhat ungrammatically[3], as [Hand, 1991] points out), rendering the "logical form" underlying (5b)[4]. This approach is however no solution to the problems of opacity and de re/ de dicto reference. More importantly, Hand [Hand, 1991] recently criticized Davidson's paratactic solution for its inability to account for negative polarity items due to the decoupling of the two parts of the logical form:

(6) (a) I didn't say that there is any beer in the fridge.
(b) I didn't say that. There is any beer in the fridge.
(c) * There is any beer in the fridge.

Hand also observes that Davidson's (demonstrative) theory of quotation could be represented paratactically, as in

(7) (a) Galileo said, "The earth moves".
(b) Galileo said$_p$ that. The earth moves.
(c) Galileo said$_s$ that. The earth moves.

where *said*$_p$ indicates quotation and *said*$_s$ indirect reported speech. Hand justifies this step pointing out that (7b), while not a form employed for direct reporting of speech,

[3]Demonstrative *that* can usually only refer to previous discourse, for following discourse *this* is the correct form.

[4]Davidson assumes a truth-functional model.

is just a small orthographic change similar to the paratactic representation for indirect reported speech:

> I see no harm in presenting Davidson's theory of direct quotation this way; the *logical form* of (7a)[5] is (7b)[6]. Correlative direct and indirect reports ... have identical logical forms, modulo the difference between $said_p$ and $said_s$.
>
> In light of this similarity, there is a curious difference between paratactic pairs involving $said_p$ and those involving $said_s$. This difference is seen in the many constructions which are allowed by direct quotation but not by indirect quotation. [p. 354]

The constructions demonstrated include topicalization, preposed directional adverbs, exclamations, etc.

Language philosophic discussions of propositional attitudes are not usually integrated into a larger model of natural language processing, but discuss one or more phenomena in isolation. In order to find a proposal how propositional attitudes interact with text representation in general, we have to look at computational models. [Nirenburg and Defrise, 1992] describe TAMERLAN, a representation scheme for texts. TAMERLAN's notion of *attitude* is one of three types of attitudes: *epistemic* attitudes refer to the epistemic status of the sentence, i.e. whether it was asserted as true, asserted as false, or qualified as known only to some degree of certainty. *Salience* refers to the theme/rheme structure of the text. An *evaluative* attitude captures the meaning of evaluative predicates, such as *favorite*. Attitudes in TAMERLAN are essentially frames that capture information that complements the propositional content of the sentence (except for *evaluative* attitudes, which represent one aspect of the propositional meaning) and attribute this information appropriately.

This notion of attitudes captures many of the ideas behind this thesis, most importantly that the aspects of a text that do not fall neatly within the propositional content still play an important role in text interpretation for real texts and therefore have to be represented appropriately.

TAMERLAN assigns numerical values to attitudes, a choice that in my opinion obfuscates the individual contribution of each attitude; only the system designers can

[5] (11) in [Hand, 1991]
[6] (12a) in [Hand, 1991]

motivate the choice of particular values and it is not obvious that the system gains anything from numbers over symbolic values. The serious shortcomings of numbers as a means of ranking attitudes is the fact that reported speech, as will be explained in detail in Chapter 3, does not represent an attitude by the reporter as much as providing evidence for the reader to develop his or her own "attitude" or evaluation of the reporetd material. This is a pragmatic inference step and should not be represented in a pure text representation, as this evaluation is likely to change from one reader to the next. Yet reported speech is similar to other attitudes in that it does not constitute a blunt assertion and leaves (more or less) room for doubt. An epistemic attitude of (a non-committal) 0.5 does not do justice to reported speech. A solution within TAMERLAN would be to add a fourth type of attitude called *evidential* which could then capture the interpreattion of reported speech developed in this thesis.

2.2.2 Speech Acts

Reported speech can be described as a *metalinguistic* device, that shows that "the grammatical language system cannot be analyzed without reference to pragmatics" [Leech, 1980, p. 31]. The field that has most systematically linked semantics and pragmatics is the speech act literature. Attitudes are at the core of speech act research, as illustrated in "Linguistic Communication and Speech Acts" [Bach and Harnish, 1979]. Bach and Harnish develop a theory of communication through speech acts. The authors consider the connection between linguistic structure and speech acts to be inferential in nature rather than purely semantic (as had been suggested previously in the literature, cf. [Searle, 1969, Sadock, 1974]).

Adopting a version of Austin's distinction between locutionary, illocutionary, and perlocutionary acts, the authors characterize the sorts of intention with which each is performed, focusing, as it were on *communicative illocutionary intentions*. These are reflexive intentions (R-intentions), which are defined in [Grice, 1957] as intended to be recognized as being intended to be recognized. Moreover, illocutionary intentions are fulfilled when they are recognized.

With this notion the authors define what they understand "communication" to mean:

> "In our view, to communicate is indeed to express a thought or, more generally, an attitude, be it a belief, an intention, a desire, or even a feeling; but in saying that to communicate is to express an attitude, we mean something very specific by "express."
>
> *Expressing:* For S to *express* an attitude is for S to R-intend the hearer to take S's utterance as reason to think S has that attitude." [Bach and Harnish, 1979, p. 15][7]

Bach and Harnish's contribution is a prediction of the path of inference that enables the communication given an utterance. This path of inference is specified in the *Speech Act Schema* (SAS). The book introduces increasingly sophisticated versions of the SAS as the authors discuss different phenomena, focusing more on the pragmatic aspects of drawing the right inferences from a given text.

Closer to the problem of reported speech, Gazdar claims that "... a theory of speech acts is the crucial ingredient of any theory of the truth conditions of utterance reports." [Gazdar, 1981, p. 65] He refers to the different truth conditions of the following sentences

(8) (a) John asserted that I would go home tomorrow.
(b) John asked me if I was going home tomorrow.
(c) John predicted I would go home tomorrow.
(d) John told me to go home tomorrow.

Answering the question of the nature of speech acts and illocutionary force, Gazdar writes: "A speech act is a function from contexts into contexts. Thus an assertion that ϕ is a function that changes a context in which the speaker is not committed to justifiable true belief in ϕ into a context in which he is so committed. ..." This quote raises two points. First, the question how the notion of commitment applies to the evidential use of utterance verbs in the context of reported speech and second, the question of evaluation of speaker beliefs by the reader.

The evidential use of utterance verbs in reported speech is a major hypothesis in this thesis. The similarity between evidentiality, i.e. the supplying of evidence for a claim, to reported speech is illustrated in Chapter 3. The reporter usually writes his or her articles based on *hearsay evidence*, obtained through interviews, memos, press releases and press conferences. The obligation to be truthful and to reveal one's evidence for every

[7]Page numbers are with respect to the second printing of the paperback edition from 1984.

claim in newspaper articles forces the reporter to reveal the source of the information transmitted. Just how important this principle is in American journalistic tradition is demonstrated by the fact that even when a source is explicitly withheld from the public, the *reliability* of the source (now of course in the eyes of the reporter) is indicated as in "From well-informed sources we hear ..." Yet if the information is attributed to someone else, the reporter does not commit to that information. The reporter is committed, however, to having characterized the circumstantial information correctly, especially if the reporter attributes a speech act to the source, he or she is committed to that interpretation. Reporters tend to be very careful not to over- or misinterpret their sources and use utterance verbs with weak semantic entailments, such as *say*, more frequently than stronger and more committing verbs, such as *promise* (which encodes the occurrence of a particular speech act in the original utterance situation). A speech act analysis thus cannot do justice to the role and semantic contribution of reported speech as it is used in newspaper articles.

2.2.3 Points of View

The second point is the question for the appropriate model of the reader of a newspaper. I claim that there cannot be one useful model for all readers in all situations. Not even, I claim, can there be a single model for an idealized automated text understander, even if the domain is limited. This topic has become known in the literature as attributing different *points of view* to an interpretation.

The sense in which I use *point of view* here has to be distinguished from the use in [Wiebe, 1990]. Wiebe is concerned with tracking the point of view from which a particular sentence, clause, or paragraph in a narrative has to be interpreted, i.e. whose subjective point of view is being expressed. For fictional texts, [Banfield, 1982] distinguishes between *objective sentences*, i.e. sentences that narrate events objectively, and *subjective sentences*, i.e. sentences that express the consciousness of an experiencing character within the story. For the interpretation of subjective sentences it is important to determine the subjective character. Point of view shifts can occur frequently in narratives and Wiebe presents an algorithm to determine the subjective character heuristically based on a model of narrative text. While propositional attitudes and reported speech (especially in form of free indirect speech) can introduce subjective

sentences and are frequently used to indicate the subjective character, this does not concern the thesis at hand. Newspaper articles do not contain subjective sentences except in the clearly marked context of a quotation and this sense of tracking point of view is therefore not important.[8]

The sense of point of view that is significant for the analysis of newspaper articles is the one implemented in ViewGen, an algorithm of belief ascription [Ballim and Wilks, 1992, Ballim *et al.*, 1991]. The authors depart from traditional work in belief logics and belief maintenance systems.

Belief logics and belief maintenance systems specify how to treat knowledge that cannot be taken for granted. This includes *maintaining* a *belief base* (BB): how to integrate a newly acquired belief, how to search the BB for a belief, and how to remove a previously held belief from the BB. The problems that arise are

Consistency: What happens when there are conflicting beliefs?

Omniscience: Is the BB closed under inference?

Negation: How is negative knowledge represented?

Retraction: How can beliefs be removed *with all their consequences*?

Other agents: How can beliefs of other agents be modeled?

Belief logics are often discussed in isolation, as an independent and consistent logic that can somehow be connected to text understanding. But often it is also assumed that a belief maintenance system is the ultimate representation system for text analysis. Consider Taylor and Whitehill:

> "Example 1: *Maggie tells Andy that she was once married.*
>
> In order to model this situation, we must make some inferences about what people believe after they say or hear something. These default rules of conversation [Bruce,

[8]This is of course not to say that newspaper articles are always written objectively. The remark concerns the newspaper style, which is objective, and not the journalists objectivity in presenting the facts. In fact Section 3.3 points out this additional level of interpretation, which lies outside linguistic analysis.

> 1978] are crucial to the construction of the belief structures. After reading the sentence in example 1, we can make the following inferences:
>
> - Maggie was married.
> - Maggie believes she was married.
> - Andy believes Maggie was married.
> - Maggie believes Andy believes Maggie was married.
> - Andy believes Maggie believes she was married.
>
> These are default inferences; some may not be made if conflicting information is already known." [Taylor and Whitehill, 1981, p. 388]

This quote exemplifies several points that hold for other models of belief as well. First, the concern is to model a *situation*, i.e. something about the real world rather than just the textual representation. Secondly, *mutual belief* is crucial to model the situation, introducing an infinite regress of belief embeddings. Thirdly, the inferences are *defeasible* and require additional inferencing to be established as beliefs themselves.

This very crude introduction of some of the problems concerning belief maintenance already shows that this field is extensive and controversial. [Ballim and Wilks, 1992] contains an excellent review of all relevant literature. The authors come to the conclusion that the areas of most active research (mutual beliefs, belief logics, etc.) are not suitable for modeling an "artificial believer" because they presuppose perfect belief partitions at the outset, assuming an omniscient algorithm. Their approach is to build a partitioned belief space based the idea of *partial* and *subjective* (i.e. limited to an individual) beliefs.

The model developed in [Ballim and Wilks, 1992] and [Ballim *et al.*, 1991], ViewGen, represents the beliefs of agents as explicit, partitioned proposition sets, called *environments*. There exist two essential types of environments, topic environments and point of view environments. Point of view environments bundle the beliefs ascribed to a particular agent, A. Nested in these point of view environments may be topic environments, bundling beliefs attributed by agent A to a particular topic, say T. Topics may be about other agents, which leads to a nested point of view environment for that agent T within the topic environment T. Consider the graphic representation in Figure 2.1. Here we can distinguish two points of view, the system's and John's, and two topics, John and Earth. Topic environments are labelled at the top, point of view environments are labelled at the bottom.

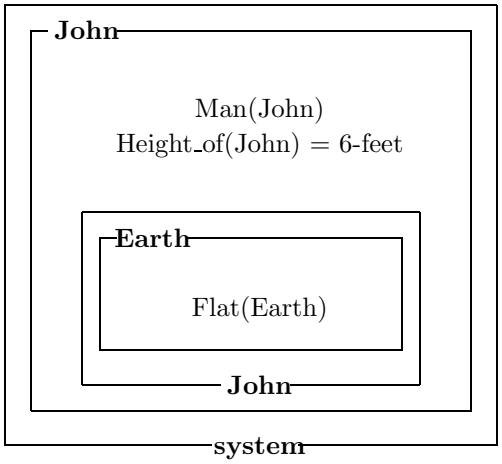


Figure 2.1: Topic and point of view environments in ViewGen

The authors present a basic default ascription rule, which is to assume that another person's view is the same as one's own except where there is explicit evidene to the contrary. This results in an amalgamation process (on demand), which ascribes beliefs from one viewpoint into another.

The authors discuss in detail how the ascription process works and how it can be used to give a uniform representation to belief ascription, metaphor resolution, and speech act resolution. To go into details would lead too far afield here; the important point is that the philosophy behind ViewGen is indeed very similar to the philosophy behind this thesis. While ViewGen provides a foundation for some of the belief literature that assumes a partitioned belief space at the outset, I view the contribution of this thesis to map out the step from text analysis to a system like ViewGen. The idea of delimited environments which can help to constrain commonsense inferencing can be seen also in Chapter 8, where I introduce a representation for reported speech that groups profiles (similar to topic environments) into *supporting groups*, thus creating a specialized form of environment that indeed can constrain the need for commonsense inferencing.

Chapter 3

Evidential Analysis

This chapter presents in detail the alternative view that the most salient aspect of reported speech in newspaper articles is that it provides *evidence for*, not *commitment to* some embedded statement[1].

As outlined in the Introduction, the *evidential analysis* for reported speech is based on the insight that reported speech is used to indicate the evidence that the reporter has for the embedded statement, which in newspaper articles usually comprises the primary information. The evidence for a statement lies in the source of the statement, which in turn determines how *reliable* the information of the statement is. Besides the identity of the source the source's expertise, insight, or relevance to the information are important factors for the determination of the source's reliability, as is the *situation* in which the statement was made. This *evaluative* information can be found in the source description and the reporting verb of the matrix clause. The evidential analysis of reported speech is in fact similar to the analysis of other *evidentials* and to motivate the evidential analysis outlined later in this chapter I will first review some of the literature on *evidentials*.

Literature on evidentials shows that language provides different devices to express the *possibility of uncertainty* of a proposition. Modals are but the crudest form to express such possibilities of uncertainty (I will in the following simply say "uncertainty" when I mean to say "possibility of uncertainty"). The uncertainty does not have to

[1]Note, that in literary texts reported speech may have quite a different role. I am also no longer including demonstrative uses of reported speech in this or the following chapters.

be the author's — often some strengthening evidence is supplied, anticipating doubts of the reader. This is especially the case in newspaper articles, where journalists write about a wide range of topics. Journalists are not usually experts on the topics they write about and the journalistic code specifies that journalists not only have to verify their sources carefully, but also have to identify them for the reader.

This "evidential" view of reported speech explains certain differences in the interpretation of reported speech in different contexts. Of special importance is the fact that the oblique context of the reported clause forces a complex pragmatic interpretation of the reported statement that can only be constrained, but not determined by semantics alone. The constraining semantic information is contained in the reporting clause; especially in the lexicalization of the *source* and the choice of *reporting verb*.

Because the interpretation of reported speech is largely pragmatic, I will limit my discussion to the special form of newspaper articles and how the context of newspaper articles influences the interpretation of reported speech (cf. Chapter 4).

3.1 Evidentials

English provides several evaluative constructs, such as *modals, negation, hypotheticals,* and *evidentials*. I claim that all these evaluative constructs require similar analysis techniques and should not be considered totally in isolation. It is, in fact, insights into the nature of *evidentials* that prompted many of the investigations I have made into the behavior of reported speech. In this section I sketch some of the literature on evidentials to illustrate its close connection to reported speech. The other evaluative constructs will not be discussed in this dissertation.

Chafe [Chafe, 1986] provides a good introduction into the phenomenon of evidentials. There he discusses evidentiality in the broadest sense, including modal auxiliaries, adverbs, and miscellaneous idiomatic phrases. Compare the following sentences:

(1) (a) It might be a spider.
(b) It is perhaps a spider.
(c) It seems to be a spider.
(d) Apparently, it is a spider.

(e) It must be a spider.
(f) I think it's a spider.
(g) It feels like a spider.

Chafe notes that the use of evidentials differs for different languages; some languages stress certain types of evidentiality over others, some languages have grammatical devices (like some American Indian languages) to express evidentiality.

Even within the English language Chafe observes different evidential behavior for written and spoken language. The difference, his study showed, lies "not so much in the frequency of evidentials overall, as in the frequency of specific kinds of evidentials." [Chafe, 1986, p.262]

Chafe develops a model for talking about evidentiality:

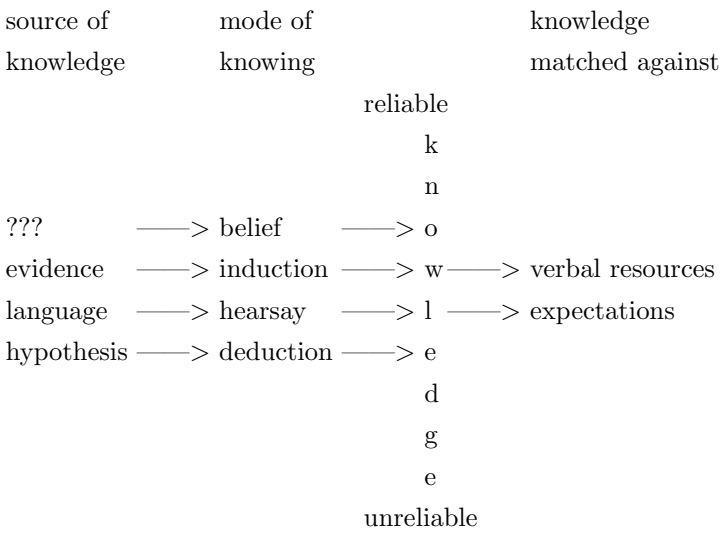


Figure 3.1: Chafe's modes of knowing.

In order to understand what evidentiality in the broad sense involves, it is useful to think in terms of several notions that are illustrated in Figure 3.1[2]. One notion can be labeled KNOWLEDGE: the basic information whose status is qualified in one way or another by markers of evidentiality. It is shown vertically in the middle of Figure 3.1. Knowledge may be regarded by a speaker (or writer) as more or less RELIABLE (or valid), as is indicated in Figure 3.1 with the suggestion of a

[2]Figure 3.1 is Figure 1 in Chafe's text.

> continuum from the most reliable knowledge, at the top, to the least reliable, at the bottom. Also relevant are various MODES OF KNOWING: various ways in which knowledge is acquired. The modes of knowing shown here are BELIEF, INDUCTION, HEARSAY, and DEDUCTION. Each of them is based on a different SOURCE, which for belief is problematic, for induction is EVIDENCE, for hearsay is LANGUAGE, and for deduction is HYPOTHESIS. (The placement of these four modes of knowing in Figure 3.1 does not imply that belief is more reliable, or deduction is less reliable, than others. Each mode of knowing can move up and down the scale of reliability.)
>
> On the right side of Figure 3.1 is an indication that knowledge may be matched against VERBAL RESOURCES (or categories), and also against EXPECTATIONS. Here there is a question of how good the match is. For one thing there is the question of whether the verbal resources which a speaker or writer chooses express more or less well the knowledge to be communicated. There is also the question of the match between a speaker or writer's knowledge and the speaker's, writer's, or other's expectations. [Chafe, 1986, pp. 262–263]

Chafe's study compared data from dinner table conversations with academic writing "highlighting", in his words, "the extremes of 'spokenness' and 'writtenness' in our data." [Chafe, 1986, p.262] He presents empirical findings for eight categories of evidentiality, namely degrees of reliability, belief, induction, sensory evidence, hearsay evidence, deduction, hedges, and expectations, giving a frequency count for each category of evidentiality for both of his corpora. The findings are summarized here in tabular form in Figure 3.2[3].

Chafe summarizes his findings as follows:

> Thus, in general, conversational English and academic writing both show a concern for the reliability of knowledge, as well as for induction. Academic writing shows more concern for deduction, neither makes a big point of marking the kind of evidence per se, and hedging as well as other devices which match knowledge against expectations are more characteristic of conversation than of academic written language. These differences can be seen as stemming from the spontaneity and interactiveness of speaking and the deliberateness and detachment of writing. [Chafe, 1986, p. 272]

One of Chafe's observations is that while English provides the language constructs to express 'evidence', English speakers rarely make use of them. This contrasts with other

[3] In academic writing *citation* is a very frequent marker of hearsay. Chafe did not include citations in this count. This explains why no hearsay evidentials where found in academic writing. Note also that "hedges" in Figure 3.2 comprise a subset of evidentials, in contrast to the use of the term in Quirk *et al*, where "hedges" included almost all forms of evidentials.

Category	**Example**	count per 1000 s	w
degrees of reliability	*This perhaps served to ward off comments from peers or teacher.*	4.6	6.8
statistical sense of reliability	*These are generally taken to be in the area of 'performance'.*	4.18	10.22
belief	*I guess I was thinking about it in a different way.*	3.6	0.6
induction	*It had evidently been under snow.*	1.3	2.4
sensory evidence	*Most of the time I feel extremely safe here.*	1.1	0
hearsay evidence	*They were using more verbs than English speaking kids have been said to learn.*	0.4	0
deduction	*Adults presumably are capable of purely logical thought.*	2.9	4.4
hedges	*. . . a Mohawk community about 30 miles from Montreal.*	3.6	0.4
expectations	*In fact this whole week has been awful.*	17	13.9

Figure 3.2: Chafe's categories of evidentiality.

languages, where the kind of evidence *has* to be specified and is sometimes obligatorily encoded grammatically (for example as suffixes).

Chafe's study did not attempt to provide a full analysis of the use of evidentials in English. Academic writing is not the place to allow doubt, which is a major function of evidentials. Other styles give a different picture. For newspaper articles, for example, we find that expressions of the category *hearsay*, which occurred hardly at all in Chafe's data, is expressed explicitly and *very* frequently in the form of reported speech. I would expect that police reports contain proportionally more evidentials of the sensory and inductive kind, but would be surprised to find a substantial count of evidentials of any kind in software manuals. Chafe's study has shown that the use of evidentials in English is a matter of style. It would be interesting to study corpora of different styles to confirm the intuitions raised by the report. The next chapter will make a first step in that direction and characterize newspaper style and the role of evidentials in news reports.

3.1.1 Evidentiality and Truth

To supply evidence for a proposition — be it to strengthen or to weaken that proposition — implies that the hearer or reader could doubt that prosition to be true. It does not necessarily mean that the speaker has any doubt in that proposition. In fact reports of sensory evidence are usually intended to *remove* all doubt (*I saw it with my own eyes ...*). In this evidentiality is crucially distinct from basic modals. Modals leave room for unquantified doubt. Evidentials constrain possible doubt and help the reader or hearer to resolve doubt *according to his or her beliefs.*

Givón [Givón, 1982] criticises the view within traditional epistemology that "the essence of sentential mode is the matter of 'truth'." [Givón, 1982, p. 24] Rather, he suggests,

> "at the bottom of propositional/sentential modalities lies the *implicit contract* between speaker and hearer, a contract specifying three types of propositions:
>
> (a) Propositions which are to be *taken for granted,* via the force of diverse *conventions* as *unchallangeable* by the hearer and thus *requiring no evidentiary justifications* by the speaker;
>
> (b) Propositions which are *asserted with relative confidence,* are *open to challenge* by the hearer and thus require — or admit — *evidentiary justification;* and finally,
>
> (c) Proposition (sic) that are *asserted with doubt,* as *hypotheses,* and are thus *beneath* both challenge and evidentiary substantiation. They are, in terms of the implicit communicative contract, 'not worth the trouble'. " [ibid]

Givón shows the encoding of evidentiality in three "typologically and genetically diverse languages" [Givón, 1982, p26] and concludes that the epistemic (or evidential) space in human language, although a graded scale, can be partitioned into three segments as follows:

Highest certainty (derived by contract) is implicit when the underlying proposition is *deictically obvious, presupposed, given by revelation, apriori-synthetic* (i.e. shared knowledge of the universe as coded in the lexicon), or *analytic* (i.e. shared knowledge of the rules of various games, including logic). In this category **evidentiality is not required**.

Medium certainty (derived by evidence) applies to *realis* assertions. In this category **evidentiality is required**.

Lowest certainty (derived by hypothesis) applies to *irrealis* assertions. In this category **evidentiality is impossible**.

I agree with Givón that a truth value cannot be the full meaning of a sentence. [Frajzyngier, 1985] attacks Givón's view, trying to prove that the indicative mood always signals truth and that in languages that have indicative mood, it is the unmarked case. Yet this criticism misses the point. It may be the meaning of the indicative mood to indicate intended truth; for reported speech this only means that the reported situation was as reported (i.e. that the source is correctly identified, that the interpretation given through the choice of the reporting verb is accurate, and that the reported clause gives an appropriate paraphrase or quote of the original utterance.) But if we look at a sentence like

"Mother said you have to go to bed."

truth is hardly what matters. It is the impact — here the force of an *order* — that is the meaning of this sentence, rather as in

"You have to go to bed (Mother said so)".

This is of course an observation similar to those that gave rise to investigations into speech acts. However, as discussed in Chapter 2, illocutionary force is not sufficient to distinguish between different reporting events. I will outline a richer representation in the following sections.

3.2 Reported Speech

The arguments presented in this section suggest that evaluative constructs are in fact a systematic, not an accidental phenomenon. I propose that all evaluative constructs should have similar interpretation and representation mechanisms; that is, negation, modality, evidentiality, and reported speech should all be treated as *metalevel comments on how to interpret the basic proposition* (or *primary information*).

3.2.1 Structure of Reported Speech

Syntax

Reported speech forms a simple syntactic paradigm, generally described by the patterns in Example (2).

Syntactic Paradigm for Reporting Verbs

(2) <Source> <reporting verb> "<utterance>"
"<Part of utterance>", <reporting verb> <source>, "<rest of utterance>"
"<Part of utterance>", <source> <reporting verb>, "<rest of utterance>"
<Source> <reporting verb> (that) <paraphrase of utterance>

The reporting verb specifies a relation between the *source* (usually the subject of the sentence) and the *utterance*; more specifically this relation is an *attribution* of the reported material to the source, similar to the notion of *belief ascription* [Ballim and Wilks, 1992].

Semantics

A reporter C uses a reporting verb to attribute an utterance B to a source A. The reporter C thereby commits to B being an accurate interpretation of the original utterance B_{org} in the original context under some aspect specified in the surrounding context. As [Clark and Gerrig, 1990] observe, a correct report does not only not have to be literal, its propositional content may be very different from the propositional content of the original utterance and can still accurately convey some aspect of the original utterance.

We can characterize the basic meaning of a reporting verb as:

(3) *paraphrase-of(B_{org},B) & utter(A,B_{org}) & utter(C, utter(A,B))*

There is a strong default assumption associated with reported speech that the reporter witnessed the source making the original utterance; if context doesn't override this assumption, the Gricean maxims [Grice, 1967] indicate that indirect evidence has to be marked in some way (see previous section). Thus we can refine (3) as:

Def. 1: The semantic field of reporting verbs has the basic semantic definition: *paraphrase-of(B_{org},B) & utter(A,B_{org}) & utter(C, utter(A,B)) & default: witness(C, utter(A,B_{org}))*

Attribution

Attributing information to a source introduces a context shift in a text. The author/speaker is *not* committed to the truth of the material attributed to someone else. Moreover, the author/speaker does not have to *believe* the attributed material. For general quotation the propositional content of the reported material does not even have to be "relevant", as [Clark and Gerrig, 1990] point out; I will argue later on that this is not the case for reported speech in newspaper articles.

The author/speaker is, however, bound to be truthful about his or her reporting. This translates into marking sufficiently what role the reported material plays in the overall current context and correctly demonstrating the relevant aspects of the report. This includes characterizing the original speech situation sufficiently for the hearer/reader to be able to evaluate it.

3.3 Evaluation of Reported Speech

Reading a newspaper is a complex activity which requires different levels of interpretation. If we assume a reader's goal to be the extraction of knowledge, we have to distinguish two cases, (i) where the reader is a "naive" reader and takes everything that is written in the newspaper at face value and moreover assumes everything that is written to be true, and (ii) where the reader has a set of beliefs and points of view that were gained from prior experiences. The second model is clearly more common, let me elaborate on this.

The "informed" reader applies background knowledge to the interpretation of a newspaper article. This background knowledge includes assumptions about the goals, interests, and biases of the newspaper as an institution. If the informed reader is also an avid reader, he or she will also have assumptions about the journalists opinions,

beliefs, points of view, and biases[4]. The nesting of assumptions about beliefs can be demonstrated using the notation of [Ballim and Wilks, 1992].

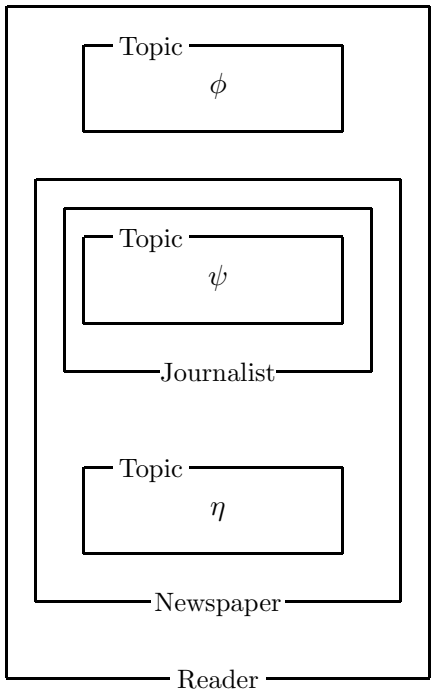


Figure 3.3: Interpretation model for newspaper articles

Figure 3.3 describes the situation. *Topic* here refers to the topic of the article, for instance "Western hostages in Lebanon". ϕ, ψ, and η refer to a (set of) belief(s) held by the respective entity. We see that the journalist environment is embedded within the newspaper environment. This embedding reflects the fact that while a journalist's expressed point of view may differ from the general point of view of the newspaper (thus $\psi \not\Leftrightarrow \eta$), the default assumption is that the views of the newspaper hold for the journalist, too. For the informed reader the task of reading a newspaper article then involves two steps, namely to interpret the article within the viewpoint of the journalist and then to assess the reader's own *meta-evaluation*, comparing the assumed points of view of the journalist and the newspaper to the reader's own.

The second step, the evaluation of the newspaper article filtered through the assessment of the newspaper's and the journalist's point of view is beyond the scope of

[4]This level is of course only relevant for articles with identified authors.

this thesis and will not be considered further.[5] It is however an important part of the motivation of the different issues raised in the following chapters and should be kept in mind.

With respect to reported speech evaluation the first step, the evaluation of the article within the viewpoint of the journalist, in itself has two substeps, as hinted at in Definition 1:

The function of the report: The *reported speech* has to be evaluated within the overall text in which it occurs, the *current context*.

The original utterance situation: The reported material has to be evaluated with respect to the situation of the original utterance, i.e. the *original context*, which is usually summarized in the *circumstantial information* of the matrix clause.[6]

The goal is to make determination of the function of the reported speech in the overall article analogous to determining the function of a simple assertion within the text. This requires the determination of the text (or discourse) structure (see [Grosz and Sidner, 1986], [Polanyi and Scha, 1984] for some approaches), making use of coherence relations (see [Hobbs, 1979, Hobbs, 1990], [Mann and Thompson, 1986] for two different notions) and possibly argument structure (see [Cohen, 1987]). This thesis will not expand on research on text structure. For the outline of a text representation in Chapters 8 and 9 I will assume a traditional model.

The important insight is that the second step, the (partial) determination of the original utterance situation, has to precede the determination of the function of the reported speech in newspaper articles. This can be done in a local context: the reporting clause provides a partial description of the original utterance situation, providing an *evaluative* environment. Within this evaluative environment the reliability of the embedded statement can be determined. Assuming that in newspaper articles the embedded statement provides the important primary information, the embedded statement

[5] Technically speaking, this step takes the propositions of the article to be, in the terminology of [Ballim and Wilks, 1992], atypical beliefs of the journalist, akin to expert knowledge.

[6] In newspaper articles the situation of the original utterance usually has to be incrementally reconstructed from the circumstantial information of several instances of reported speech pertaining to the same original utterance situation. For the evaluation of a particular instance of reported speech, however, the information preceding the occurrence is sufficient.

can now be treated analogously to a simple assertion in the determination of text structure, provided that the evaluative environment has been captured appropriately. This is just what the evidential analysis of reported speech does. The following sections will introduce the important role that lexical semantics plays in the analysis of the evaluative environment of the reporting clause and then give a formalization of the evidential analysis itself.

3.3.1 Evaluative context of the reported clause

Let us first explore what role lexical semantics plays in the second evaluation task, the interpretation of the original utterance. The matrix clause carries explicit information about the situation of the original utterance. The matrix clause has two major components, the reporting verb and the source, which I will consider separately.

Lexical Semantics of the Reporting Verb

Reporting verbs partly specify the situation of the original utterance of the reported material as to manner, intention, force, etc. Modifications by means of prepositional phrases, adjuncts, adjectives, or adverbs can further specify the situation of the original utterance; I will, however, not consider these constructions here.

vocalization	neutral	contextual	pragmatic intent
cry	say	allude	accuse
call	tell	argue	admit
mumble	report	contend	claim
mutter	relate	insist	deny
scream	announce	reiterate	promise
shout	release	reply	joke
stammer	state	dispute	assure
stutter	add	ask	pledge

Figure 3.4: Reporting verbs encoding the dimension *manner*

The field of reporting verbs is differentiated along several *semantic dimensions*. Fig-

ure 3.4 shows a selection of reporting verbs and their general contribution to reconstruct the original situation. Here the classification criterion is *manner*. *Neutral* lists verbs that do not specify the manner of the original utterance; *vocalization* specializes physical aspects of the original utterance such as low or high voice. *Contextual* manner refers to the coherence of the original utterance in its original context; it specifies, for example, whether the statement was a reiteration of a previous point or whether it was an allusion. *Pragmatic intent* specifies, beyond the textual manner, an intention with pragmatic ramifications, such as a promise, a pledge, etc.

Another important dimension of classification is the textual status of the reported material within the context of the original utterance. Here we distinguish *new* and *previously mentioned or implied* material. Within the class of previously mentioned or implied material we can further distinguish *confirming* material from *contrasting* material. Some examples are presented in Figure 3.5.

new	previously mentioned or implied	
	confirming	contrasting
announce	agree	deny
report	confirm	insist
tell	consent	refute

Figure 3.5: Reporting verbs encoding the dimension *textual status*.

Figures 3.4 and 3.5 show only two possible semantic dimensions of the field of reporting verbs. They are meant to serve as examples here. It is important to analyze the semantic dimensions of a semantic field before attempting to define the lexical semantics of the individual members of the field. Only when the field is considered in its entirety can arbitrariness in the lexical definitions be largely avoided[7].

A contrasting study of the semantic dimensions used by two dictionaries can be found in Chapter 7.

Semantic dimensions describe the implicit structure of a semantic field. Only knowledge of all dimensions relevant to the field, even those that are not part of the meaning

[7]This is not the current practice of lexicographers when writing entries for standard dictionaries. This might exlpain certain inconsistencies common in dictionaries.

of an individual word. Consider the reporting verb sense of *say*, the most unmarked reporting verb. *Say* is defined partly through the absence of any semantic dimension in its definition, which is also what distinguishes *say* from other words in the field.

The Lexicalization of the Source

Equally important to the task of evaluating the reported speech is the lexicalization of the source. While the *identity* of the source already gives important information about the situation of the original utterance, the way the reporter refers to that source, the *lexical realization*, indicates, how the reporter wants the reader to interpret the reported material. It is thus part of both levels of interpretation.

As mentioned in Section 3.1.1, the main issue in interpreting the original utterance is the determination of its *reliability*[8]. The assessment of the reliability is directly dependent on the assessment of the source. We can distinguish two major reasons for trustworthiness: *inherent* and *incidental*. Inherent trustworthiness stems from the *authority* or the *expertise* of the source; incidental trustworthiness stems from the evidence available to the source or from the involvement of the source in events related to the reported material. Examples are:

Authority: *official*

Expertise: *researchers*

Evidence: *witness*

Involvement: *victim*

These categories are not mutually exclusive; a *specialist* has expertise in his/her field but also most likely the best evidence. These examples further demonstrate that trustworthiness is usually not assigned to a person in general — even the most trusted person may not have sufficient insight into certain matters. That is, trustworthiness is context dependent. For reported speech, trustworthiness is relative to the reported material: an *expert* is trusted in his/her domain; *parents* are trustworthy on childcare

[8] *Reliability* here also stands for *certainty* and *credibility*.

and schooling matters; the *President of the United States* has authority on foreign policy affairs; and the *President of the NFL* on questions regarding the football season. To establish the competency of the source is a pragmatic issue I will largely ignore here. For newspaper articles, however, this is feasible since there is a strong assumption that the reporter has selected competent sources. Nevertheless, unless the person's name is generally known, reporters usually indicate the competence of their sources through the choice of *lexical realization*. The task is to determine how the lexical semantics combines to characterize the source. Examples (4) show initial descriptions of sources in the Wall Street Journal corpus.

(4) (a) A federal judge
(b) Big Board Chairman John Phelan
(c) Economics Minister Helmut Haussmann
(d) Edward Egnuss, a White Plains, N.Y., investor and electronics sales executive,
(e) F.H. Faulding & Co., an Australian pharmaceuticals company,
(f) New Brunswick Scientific Co., a maker of biotechnology instrumentation and equipment,
(g) Prime Minister Lee Kuan Yew, Singapore's leader and one of Asia's leading statesmen for 30 years,
(h) Robert L. Bernstein, chairman and president of Random House Inc.,
(i) Two leading constitutional-law experts
(j) a federal judge
(k) definitive figures from the National Statistics Institute

Source NPs are very regular, so regular in fact, that a partial semantic grammar can be provided based on the analysis of a 250,000 word corpus of TIME Magazine articles (from 1963). The analysis of the subjects for all occurrences of seven reporting verbs in the corpus yielded a limited number of different types of lexicalizations of the subject. Interestingly, this grammar shows that the source is often lexcalized as some *institution* or as holding a particular *position*, indicating the tendency for reporting verbs to accomodate metonymy in subject position (cf. Section 6.4.2).

source ⟶
[quant] [mod] descriptor [“,” name “,”] |
[descriptor | ((a | the) mod)] [mod] name |
[inst ’s | name ’s] descriptor [name] |
name “,” [a | the] [relation prep] descriptor |
name “,” [a | the] name ’s (descriptor | relation) |

name "," free relative clause
descriptor ⟶
role |
[inst] position |
[position (for | of)] [quant] inst name ⟶
[title] [first] [middle] last ["," spec] ["," age ","]
position ⟶
minister | *official* | *chief* | *president* | *neurosurgeon* | ...
inst ⟶
[location] [allegiance] [institution]
institution ⟶
U.N. | *police* | *governement* | *city water authority* | ...
location ⟶
country | *city* | *state* | *region*
allegiance ⟶
political party | *religion* | ...

A partial semantic grammar for source NPs

The grammar exemplified is partial — it only captures the regularities found for a small set of verbs in the TIMEcorpus. Source NPs, like all NPs, can be adorned with modifiers, temporal adjuncts, appositions, and relative clauses of any shape. The important observation is that these cases are very rare in the corpus data and must be dealt with by general (i.e. syntactic) principles.

The value of a specialized semantic grammar for source NPs is that it provides a powerful interface between lexical semantics, syntax, and compositional semantics. The source NP grammar compiles different kinds of knowledge. It spells out explicitly that logical metonymy is to be expected in the context of reporting verbs. Moreover, it *restricts* possible metonymies: the *ham sandwich* is not a typical source with reporting verbs. The source grammar also gives a likely *ordering* of pertinent information as roughly

COUNTRY|LOCATION ALLEGIANCE INSTITUTION POSITION NAME.

This information defines essentially the *schema* for the representation of the source in the knowledge extraction domain. The usefulness of semantic grammars for robust parsing has recently been demonstrated by [McDonald, 1991] and is currently under

further development in the TIPSTER project at Brandeis.

In general, source NPs specify one or more of the following:

- Name of the source
- Group, company, or institution, for which the source stands as a representative or spokesperson
- Profession or field of expertise of the source
- Purpose, product, or service provided by the group, company, or institution the source represents
- Location of the source
- Political, religious, or other allegiances relevant to the topic

3.3.2 Evaluative context of the reporting clause

The evaluative context of the reporting clause is the entire text and thus difficult to characterize in a general way. I will limit the discussion here to some factors that directly influence the interpretation of the reported clause as well.

The distinction is one of function. The evaluative context of the reporting clause, or current context for short, has to determine whether the propositional content of the reported clause is to be considered *primary information* or *ancillary information*. The distinction is important to direct the reasoning process that attempts to construct coherence relations. *Ancillary information* comes from reported material that is not relevant to the argument or story being developed, but that serves to give some support to a point that has already been made explicitly or that is underlying implicitly. Figure 3.6 illustrates the point.

C3 is an instance of ancillary information. Rather than contributing to the topic, C3 elaborates on the original context, illustrating Fitzwater's attitude and the level of formality and R3 further specifies the context as a press conference. C4 can be seen as ancillary information as well.

US Advising Third Parties on Hostages

(R1) *The Bush administration continued to insist yesterday that* (C1) it is not involved in negotiations over the Western hostages in Lebanon, (R2) *but acknowledged that* (C2) US officials have provided advice to and have been kept informed by "people at all levels" who are holding such talks.

(C3) "There's a lot happening, and I don't want to be discouraging," (R3) *Marlin Fitzwater, the president's spokesman, told reporters.* (R4) *But Fitzwater stressed that* (C4) he was not trying to fuel speculation about any impending release, (R5) *and said* (C5) there was "no reason to believe" the situation had changed.

(A1) Nevertheless, it appears that it has. ...

Figure 3.6: Boston Globe, March 6, 1990

Primary information, on the other hand, is information that pushes the point, argument, story, or narrative forward. C1, C2, and C5 are clear examples of primary information.

The distinction between primary and ancillary information can only be made when the larger context of the text is considered. It is the text structure in particular that determines it.

3.4 Evidential analysis of reported speech

The evidential analysis of reported speech in newspaper articles assumes that the reported clause contains primary information which the reporter wants to convey to the reader, but which is "uncertain" to the degree that the reporter has no first hand knowledge or expertise and is attributing the information to a source. This does not imply that the reporter does not have an opinion of his or her own; to the contrary, the role that the reported speech takes in the article indicates the reporter's evaluation of it (compare the text in Figure 3.6, where the beginning of the second paragraph more or less implies that Fitzwater's statements are obfuscating the matter). It is newspaper convention to "prove" one's point by citing a source for the strength of the argument (newspaper stories fall into the second of Givón's categories, that is they need evidential support.) This allows the reader to follow the reasoning process and to go along or differ

based on reasonable grounds.

The reporter will characterize the reported material not only through the overall argument presented, but also through the lexical choice for reporting verb and the description of the source. The lexicalization of the reporting verb and source together can either indicate high or low authority. Thus the evaluation of the reported speech spans from the "low" end of lexical semantics all the way to the "high" end of the whole (argumentative) structure of the text.

We can summarize the findings of this chapter so far in Definition 2:

> **Def. 2:** In the evidential analysis of reported speech the reported clause is primary information, assumed to be true with the certainty indicated by specified context variables.
> The context variables that determine the certainty of a reported clause are:
> - the source,
> - the context of the original utterance (original context) as encoded in the reporting verb and possible modifications found in the matrix clause, and
> - the current context of the reported speech, as encoded in the argumentative or text structure.

I will call the fact that the primary information of the reported clause is valid to the extent that the evidential support warrants this its *evidential scope*. I propose the following notation for evidential scope:

S'[OC,CC,TC]

where

S' is the interpretation of the (reported) clause,
OC is a variable for the original context,
CC is a variable for the current context, and
TC is a variable for the temporal context.

OC, CC, and TC are *context variables*, which indicate the original context (OC), as encoded in the reporting clause, the current context (CC), that is the position in the text structure or argument structure of the text, and the temporal context (TC), that is the position in the trace.[9]

[9]The notation is not only applicable to reported speech contexts and other constructs spanning orthogonal contexts, but can also be useful for simple sentences, where the OC could contain the reference for anaphora etc.

Evidential scope is not limited to reported speech. It occurs, in fact, whenever a sentence introduces a context that is orthogonal to the current (textual) context, as do hypotheticals and verbs of cognition (*John seeks a unicorn*). This is not a novel observation; in model-theoretic semantics, for example, this problem has been addressed by possible worlds semantics [Hintikka, 1969] and by Discourse Representation Theory (DRT) [Kamp, 1988, Kamp and Reyle, 1991].

The analysis proposed in DRT [Kamp and Reyle, 1991] is similar in spirit. DRT notation is focused on making explicit the scope of discourse variables etc. in the form of nested boxes. The notation allows to determine quantifier scope, anaphora, temporal relationships etc. DRT notation is rather complex and it goes beyond the scope of this thesis to introduce the theory in detail. However, DRT is one of the best developed formal semantic theories that addresses complex issues such as complement sentences.

In [Kamp and Reyle, 1991] one section is devoted to that-complements. It suggests an analysis that shows similarities to the analysis presented here. In order to compare the two approaches, some basic notions in DRT have to be introduced.

3.4.1 Discourse Representation Theory

Discourse Representation Theory, DRT, is a formal semantics that focuses on representing complex discourse. Thus the aim of a discourse representation structure (DRS), the representational unit in DRT, is to make explicit the interconnection between clauses in a larger text. Much work has been done on temporal relationships between events (cf. [Reyle, 1986]), on conditional sentences (cf. [Kamp and Reyle, 1991][10]), anaphora resolution (ibid.), and conjunction (ibid.). I will not discuss the semantics underlying DRT (and how that semantics is linked to the syntactic analysis) but simply explain the notation to a level of detail that allows me to introduce DRT's solution to that-complement sentences.

The basic building block of a discourse representation structure is a box which delimits the scope of the formal variables introduced inside. Formal variables are introduced for every possible discourse referent. Propositions are represented as propositions over

[10] [Kamp and Reyle, 1991] is the newest text on DRT and gives a good overview over previous research, integrating it into a very readable outline of the whole enterprise.

formal variables. Thus a simple sentence (5) is represented as DRS (6):[11]

(5) Jones owns a Porsche.

(6)

x y
Jones(x) Porsche(y) x owns y

Anaphora are represented by unifying formal variables.

(7) Jones owns a Porsche. It fascinates him.

(8)

x y u v
Jones(x) Porsche(y) x owns y u = x v = y v fascinates u

DRSs can be nested to delimit the scope of certain operators. Negation, for instance, is represented as applying to an embedded DRS:

(9) Jones owns a Porsche. He does not like it.

(10)

x y u v
Jones(x) Porsche(y) x owns y u = y v = x ¬ [v likes u]

[11]The following examples are all from [Kamp and Reyle, 1991] or slight adaptations thereof.

With this brief introduction, let me discuss the example given for intensional contexts. I include here some of the original discussion of [Kamp and Reyle, 1991] who make their point most succinctly.

> Consider the following two sentences text (free after Stalnaker: Belief Attribution and Context)
>
> (I) Phoebe believes that a man entered her garden yesterday. She believes he stole her prize zucchini.
>
> The first two steps in the DRS construction for the initial sentence of (I) yield:

(II)

n s x
Phoebe(x)
n ⊆ s
s: [x believe that a man entered her garden yesterday]

> The next step involves a non-anaphoric application of the new rule:

(III)

n s x p
Phoebe(x)
n ⊆ s
s: [x believe p]
{ p : [a man entered her garden yesterday] } K

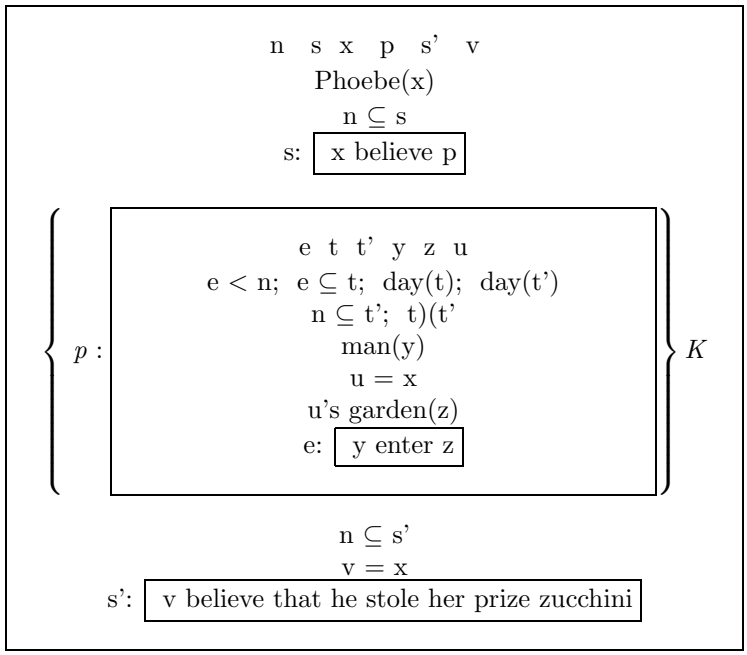
> Processing the complement sentence inside the delineated DRS K, ..., yields the structure:

(IV)

n s x p
Phoebe(x)
n ⊆ s
s: [x believe p]
{ p : [e t t' y z u ; e < n; e ⊆ t; day(t); day(t'); n ⊆ t'; t)(t'; man(y); u = x; u's garden(z); e: [y enter z]] } K

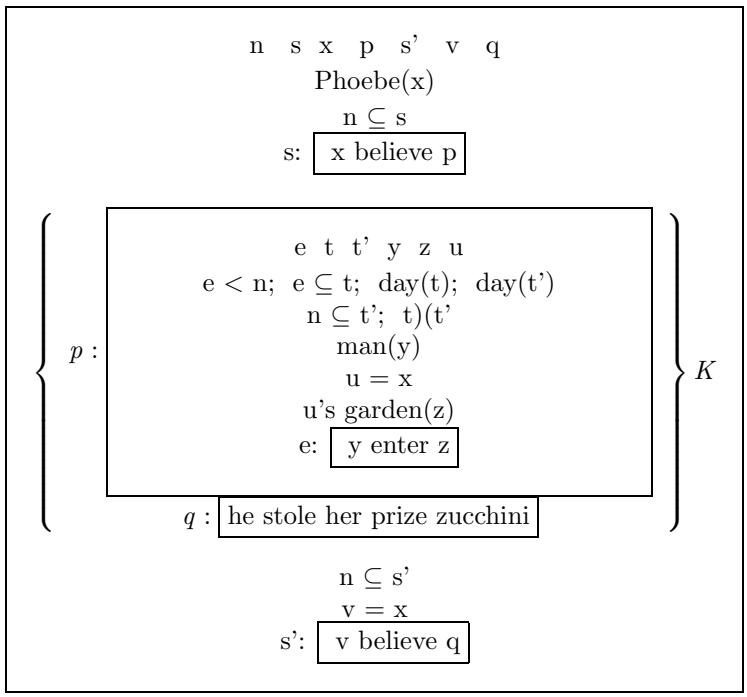
This completes the DRS for the first sentence of (1). The first two steps in the processing of the second sentence extend (4) to

(V)



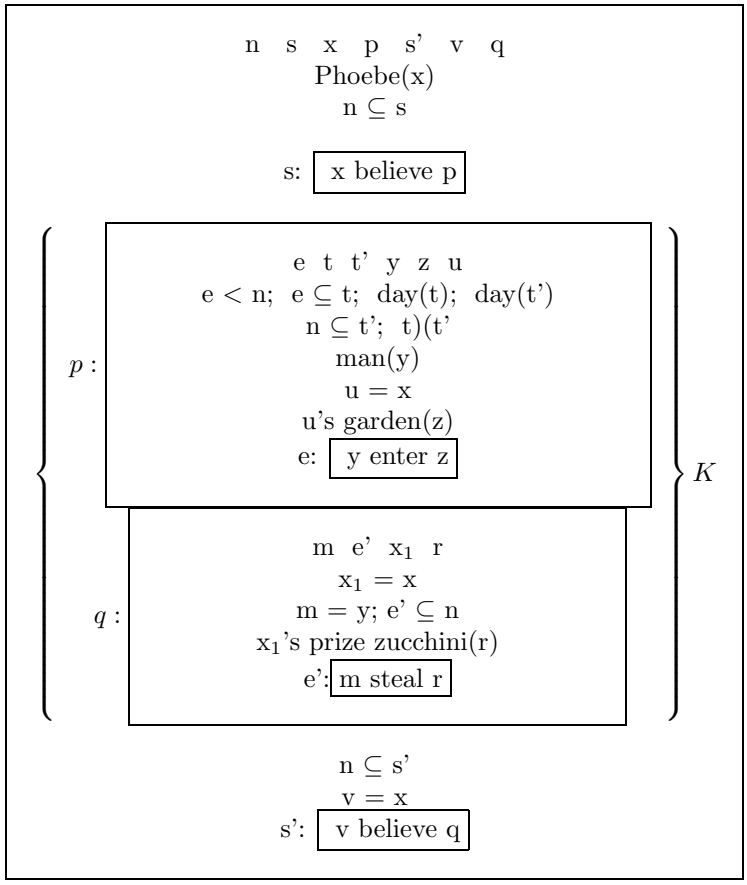
The next step involves another application of the new rule. But this time it is an anaphoric, not a non-anaphoric application. This means that the expression q:$\langle\phi$, {he stole her prize zucchini}$\rangle$ is not made into a new delineated DRS but is added to the delineated DRS K. The result is:

(VI)



The remaining steps concern the processing of the complement sentence *he stole her prize zucchini*. Here it is of interest to note that, according to our addendum to the definition of accessibility, the intuitively intended antecedents of *he* and *her* are indeed accessible, y is accessible because it belongs to the amalgam of a suitable delineated DRS. This is the delineated DRS K, whose suitability derives from the fact that it contains a condition of the form P:K", where p also figures in the condition s: [x believe p] which stands in the right relation to the condition s': [v believe q]. x is also accessible, because it belongs to the universe of the DRS which, in the rule description above, is referred to as K'. . . . we obtain, as a result of carrying out these remaining steps, the DRS

(VII)



The interesting parallel between the DRT representation and evidential scope is the fact that DRT separates the matrix and the complement and represents the complement independently, yet indexed by the matrix clause. This suggests implicitly that evidential scope has been recognized as an issue. However, DRT does not distinguish between different *that* complement clauses and treats *believe* and *say* in the same way, assuming an intensional context for the complement. Thus the representation keeps

the complement imbedded in the complex context in which it occurred and requires a possibly unwieldy semantic interpretation procedure in order to interpret the embedded DRS as primary information for an evidential analysis. As far as the theory of DRT is detailed in this book, such a semantic interpretation procedure goes against the grain of the model developed.

Two points are of importance then about the DRT analysis of that-complements: DRT has developed the basic tools that would allow an evidential analysis for reported speech in newspaper articles. This is an important plus for the evidential analysis of reported speech proposed here. On the other hand the simple example outlined above already shows a major shortcoming of DRT as a representation formalism for information retrieval. DRT represents a formal semantics and requires a full analysis of the text before it can be queried about any part. But DRSs contain information at all levels, temporal, anaphora, conditionals, intensional contexts, and quantifiers and are not complete before all this information has been properly analyzed; this leads to a brittle system. As I will outline in Chapter 8, a distributed representation permits us to keep the temporal representation separate from the text structure, which can be separated from propositional information about discourse entities. The advantage of such a distributed representation is that analysis of one aspect of the text can proceed and be exploited before the other aspects have been fully analyzed, permitting a delayed evaluation protocol for certain text phenomena. It is important to keep in mind, however, that such a distributed representation may well be equivalent to a formal semantics such as DRT, benefitting from the insights gained therein.

DRT representation suffers from the mathematically based strictness that renders the resulting representation almost unreadable even for simple sentences. The important similarity here is that the content of the complement clause is represented independently from, yet indexed by the matrix clause, allowing in principle an evidential analysis of reported speech.

The nature of reported speech to introduce a context split is reminiscent of multiple contexts in evolving spoken discourse, where the following text can proceed in or return to either context[12]. This implies for the computational representation of the different

[12]See [Bernth, 1990] for examples of anaphoric reference into and out of an intensional context.

contexts that they both have to have similar status[13] for reference resolution, text or argumentative structure building, and summarization purposes.

In summary, we have been able to distinguish and categorize the information found in the matrix clause and in the embedded clause of reported speech. But while we know that in the case of evidential analysis the primary information resides in the embedded clause, we cannot separate the two clauses in a meaningful way. The context variables introduced above preserve all relevant context information while granting the embedded clause an independent status according to the evidential analysis. Context variables are not necessary for the understanding of the sentence in isolation. For text understanding, however, their full analysis is required.

The impact of the evidential analysis is immediately obvious when we contrast the inferential value of three different analyses of reported speech. As mentioned briefly, literature on quotation characterizes the difference between a quotation and and other sentential constituants as one of "use" vs. "mention". Quotations (and often reported speech in general) is seen as mentioning a statement and thus not making any commitments to its validity. Let me call this analysis the "mention" analysis.

A second analysis is that of [Clark and Gerrig, 1990], which I will call here the "demonstrative" analysis. Clark and Gerrig describe that a demonstrative use of reported speech can be successful even when the reported material is not repeated literally — rather than the propositional content of the original utterance situation, the demonstrative use can "demonstrate" a wide variety of *aspects*[14], such as pitch, voice, dialect, speech pattern, phrasing, mood, etc. In this analysis a spoken report of a long sermon can felicitously be rendered as "Blahblahblahblahblah..." in the right context and with the appropriate intonation etc.

Figure 3.7 contrasts the inferences that the three different analyses of reported speech license with respect to the validity of the reported material. While the "mention" analysis does not give any indication on the validity of the complement, the demonstrative analysis implies that some contextually marked aspect of the complement (and its rendition) is true. The evidential analysis in contrast makes explicit that the complement

[13] In Scheme parlance they both have to be first class objects.

[14] The aspects that are highlighted — or reported — have to be marked clearly in the immediate context.

"Mention" Analysis

SAY(Source, ϕ') is true $\not\Rightarrow$ ϕ' is true.

Demonstrative Analysis

SAY(Source, ϕ') is true $\Rightarrow$ some contextually marked aspect of ϕ is true.

Evidential Analysis

SAY(Source, ϕ') is true $\Rightarrow$ ϕ'[OC,CC,TC] is true.

Figure 3.7: Inferential impact of the evidential analysis of reported speech.

is valid to the extend that the reader judges the context variables reliable. The important consequence of the evidential analysis then is that the complement clause can be considered in isolation — as might occur for instance for information retrieval based on keyword search etc. — and that it can be evaluated on demand in a localized context. This context is preserved in the context variables that indicate reliability. Additional context can be provided using the text representation scheme developed in Chapters 8 and 9, where the concepts of *profile* and *supporting group* give additional information for the quick and partial evaluation of a statement in isolation, enabling statistical and "intelligent" information retrieval methods to cooperate.

3.5 Example Analysis

The philosophical impact of the evidential analysis lies in the contrast to an intensional analysis. Where in an intensional analysis the truth value of the complement could not be decided at all, the evidential analysis claims that the complement can be considered true if the circumstantial information is deemed reliable. This means that full commonsense reasoning can take place over the complement clauses provided that the reliability indicating information (the circumstantial information) is retained (in case the reliability of this information is changed at a later point in time).

The practical impact of the evidential analysis as presented here is that it provides a general purpose structure, or a linguistically based indexing of the function of the components found in the matrix clause of reported speech. This explicit indexing of the function of the components makes a delayed evaluation strategy possible, where

the reliability is computed on demand, under the particular set of beliefs and points of view of the user at the time. Thus the evidential analysis requires a fair amount of overhead in preprocessing time. This overhead, I content, can be amortized amply by the reusability of the resulting representation and the savings in search time gained by the indexing.

The following discussion intends to show possible savings in search time given an evidential analysis. The examples are contrived for the purposes of showing possible advantages of this scheme for information retrieval without developing a full information retrieval system. Certain representational choices may appear arbitrary and will in a real system depend on overall constraints. They are not relevant to the main point made here. This point in short is that an evidential analysis provides a structure of the information found that allows statistical methods, such as key word searches, to yield good results in significantly reduced search time.[15] Thus, not only does the evidential analysis provide a general purpose representation that can be evaluated by systems with vastly differing belief sets, but it can also be exploited by a variety of retrieval strategies.

Let me illustrate the impact of the evidential analysis of reported speech on an analysis of the text given in Figure 3.6, page 51.

Assume that C3 and C4 have been recognized as ancillary information and will be represented separately.[16] The primary information can then be represented as follows:

```
BA   = ‘‘the Bush administration’’
E1   = λx:  NEGOTIATE(x, RELEASE(CAPTORS, ‘‘the Western hostages in Lebanon’’))
C1   = NOT(E1(BA))
R1   = (OC: (S: BA; U: REPEAT(INSIST); C: YESTERDAY),
        CC: NIL,
        TC: PAST-TENSE(R1), PRESENT-TENSE(C1))
S1   = C1(R1)
```

[15] This is a worst case analysis and does not imply that “intelligent” methods used for further processing would not yield markedly better results.

[16] Because ancillary information by definition does not push the storyline forward, its function within the text is more complex. In this case, C3 and C4 provide information about the utterance situation: Fitzwater clearly indicates that he will not commit to any evaluation of the situation except for the denial of any direct involvement of the government in negotiations. The inferences necessary to evaluate ancillary information are complex and go beyond the purpose of this thesis.

```
USO  = ‘‘US officials’’
PL   = ‘‘people at all levels’’ & E1(PL)
C2   = ADVISE(USO, PL) & INFORM(PL, USO)
R2   = (OC: (S: BA; U: ACKNOWLEDGE),
        CC: S1 ‘‘but’’ S2,
        TC: PAST-TENSE(R2), PRESENT-PERFECT(C2))
S2   = C2(R2)
MF   = ‘‘Marlin Fitzwater’’
E2   = λu:  BELIEVE(u, CHANGED(SITUATION))
C5   = ∀y:  NOT(∃z:  REASON(z, E2(y)))
R5   = (OC: (S: MF; U: SAY),
        CC: S4 ‘‘and’’ S5,
        TC: PAST-TENSE(R5), PAST-TENSE(C5))
S5   = C5(R5)
```

The variables within the original context variable OC, S, U, and C, refer to the source, reporting verb, and additional information about the original context respectively.

The example representation uses a straightforward predicate argument structure to illustrate representation of evidential scope. The representation of the propositional content, however, is not pertinent to this issue and can be replaced by a more refined representation.

The evidential analysis provides an explicit indexing of different information according to its function: sources are identified as such, primary information is identified, attitudes encoded in the reporting verb are identified. Consider for instance an implementation where “sentences”[17] are represented as frames in the following format.

SENTENCE
  PRIMARY INFORMATION: < filled with C-clauses above >
  CIRCUMSTANTIAL INFORMATION < corresponds to R-clauses above >
    OC
      S: < filled with source information >

[17]The term *sentence* in this example refers to one matrix clause with all dependent clauses and does not necessarily coincide with the sentence boundaries marked by periods in the text. Two conjoined matrix clauses would be considered two sentences. The word *sentence* is chosen because the word *clause* could be confused with the complement clause and *phrase* would be misleading also.

U: < filled with reporting verb information >
C: < filled with additional circumstantial information >
CC: < filled with text structure information >
TC
R: < filled with temporal information for matrix clause >
C: < filled with temporal information for complement >

The colons indicate slots that are to be filled from the text.

This representation can obviously make use of information from selected slots without considering all provided information. For instance if statements from a particular source are queried, the search can be limited to the subslot named S. The advantage of the evidential analysis for retrieval searches then is similar to the advantage that template representations give in classifying information in different slots. Template representations suffer, however, from loss of information that is pertinent but does not fit in a slot, from narrow specialization to one domain (making reuse for different domains impossible), and from the difficulty to add or modify slot fillers in light of new information. The evidential analysis only indexes part of the information in named slots and so avoids this problem.

A representation of an evidential analysis together with a theory of the function of reported speech in newspaper articles can be exploited in several ways for retrieval with statistical or "intelligent" methods. The following three scenarios suggest how the search for particular information can be limited to a much smaller subset of text when making use of the context variables. Note that all scenarios assume a preanalyzed text database that has not been indexed for topic areas.

Scenario 1:
Task: Retrieve all statements that the President of the U.S. made concerning the Western hostages in Lebanon in March 1990.
Comment: This scenario shows that retrieving already indexed information first (here the source information) cuts down on search time.
Approach: Restrict the articles searched to 1990. The pertinent information can be found in the dateline (omitted in the example above). Because the search is for statements by a particular entity, the U.S. govenment, it can be restricted to the evaluative environments, in particular to the slotfillers for the subslot S in the slot OC. In those environments a keyword search for all popular paraphrases for "U.S. government", such

as "the Bush administration", "the President", "the U.S.", "Fitzwater", etc. would be sufficient. If a list of paraphrases is not available, the source NPs have to be analyzed linguistically from first principles. For the final analysis the complement clauses indexed by the previously retrieved sources have to be analyzed for the topic "Western hostages in Lebanon". This particular term is a phrase that was commonly used for the hostages; a keyword search for the string and substrings ("hostages", "Lebanon") could be successful. In the more general case, however, an "intelligent" retrieval method would have to be used.

Procedure:[18]

∀ articles do
    if date ⊂ 1990
    then ∀ i do
        if Ri:OC:S ∈ paraphrase of key
        then if topic of Ci = Western hostages in Lebanon
            then retrieve article

Gain: This procedure may appear unusual, because it searches for the (much greater) mass of statements by the U.S. government and searches for the topic within these statements rather than selecting first the articles of the relevant topic and searching for U.S. government statements within those. But in a text base organized according to an evidential analysis of the texts, selecting statements made by the U.S. government is a trivial task compared to searching for a topic area, because sources are indexed in specially marked source subslots of the OC variable. Note that this approach is particularly beneficial in connection with massively parallel architectures, where indexed primary information can be stored at the processor level and retrieving a particular source can be very fast.

Scenario 2:
Task: Retrieve all articles that concern the hostages held in Lebanon in 1990.
Comment: This scenario shows that knowledge about the style of reported speech in

[18]This procedure suggests that articles are analyzed one by one whether they contain a source corresponding to the U.S. government and if so to analyze the topic of the statement. This is merely a notational convenience. In a realistic system some further indexing will for example provide for a list of articles from 1990. This set can then be filtered for articles containing statements by the U.S. government. This resulting set of articles can then be analyzed for the topic. The intermediate sets of articles can be retained and appropriately indexed speeding up retrieval of this subset for the next appropriate query.

newspaper articles in connection with an evidential analysis can cut down the search space even for broadly specified topics. The insight here is that the topic can be found analyzing only the primary information. This is a heuristic and might fail in certain (rare) cases.

Approach: Restrict search to articles from 1990. Restrict search to primary information (i.e. the C-clauses in the text above). Search for a list of keys, such as "hostage(s)", "Western hostages", "Lebanon", names of hostages, etc. or perform "intelligent" retrieval.

Procedure:

∀ articles do
    if date ⊂ 1990
    then ∀ i do
        if topic of Ci = Western hostages in Lebanon
        then retrieve article

Gain: The gain in this procedure is less obvious, because the topic has to be determined for the large amount of articles printed in 1990. However the search here is limited to the primary information of the complement clauses, cutting the search space down significantly.

Scenario 3:

Task: Retrieve articles that show U.S. government involvement in negotiations over Western hostages in Lebanon in March 1990.

Comment: This query shows that reliability information is crucial in improving preciseness of recall.

Approach: Restrict search to articles from March 1990. Restrict search to articles concerning Western hostages in Lebanon as in scenario 2. Analyze complements (C-clauses) for "negotiations" and paraphrases of "possible release". For every statement thus obtained, evaluate reliability by assessing the reader's trust in the source and combining the result with the reliability implicit in the lexical semantics of the reporting verb. If the reliability satisfies a certain criterion (i.e. is "high", as in either passing a threshold or being among the n most reliable found in the text base), retrieve the article.

Procedure:

∀ articles do
    if date ⊂ March 1990
    then ∀ i do

if topic of Ci = U.S. government involvement in negotiations
over Western hostages in Lebanon
then if RELIABILITY(Ri) = high
then retrieve article

where RELIABILITY is illustrated for the first two sentences in out text as follows:

RELIABILITY(R1)
COMBINE
RELIABILITY(R1:OC:S) < mass term, no responsible source specified >
RELIABILITY(R1:OC:U) < *insist*, presupposes opposition >
RELIABILITY(R1:OC:C) < no filler >

RELIABILITY(R2)
COMBINE
RELIABILITY(R2:OC:S) < mass term, no responsible source specified >
RELIABILITY(R2:OC:U) < *acknowledge*, factive verb >
RELIABILITY(R2:OC:C) < no filler >

Gain: The major work for the preciseness of retrieval of articles in this procedure is done by the subprocedure RELIABILITY. Even though an evaluation criterion has not been given here, it should be obvious that the higher strength of R2 should result in retrieval of the article analyzed above, even though the Bush administration denied involvement in S1. This is an indirect analysis of the notion of "negotiation", which an intelligent system could equate with "providing advice to and being informed by . . . ", the content of C2.

The three scenarios provide only a very rough outline of the usefulness of the evidential analysis combined with a thorough understanding of the role of reported speech in newspaper articles. Some steps outlined here will become clearer as I address issues of newspaper style, lexical semantics, and argumentative structure in the following chapters. Better tools still will be described in Chapter 8, where *profiles* and *supporting groups* are introduced.

Chapter 4

Newspaper Style

The previous chapter outlined the evidential analysis of reported speech, how it compares to some literature on evidentials and quotation, and how it could be useful for information retrieval purposes. To perform an evidential analysis means to preserve the character of the content of the matrix clause as *evidence* for the reliability of the complement, that is to preserve the information for later evaluation by a reader or user program. This delayed evaluation scheme permits different interpretations to be gained from the same text based on different belief bases, different tasks, or even differing new information at a later point in time.

The evaluation of the reliability is expressly *not* part of the evidential analysis itself. Yet it is important to motivate the usefulness of an evidential analysis by showing what additional linguistic and conventionalized constraints there are that facilitate the evaluation process.

The next three chapters analyze different phenomena, newspaper style and lexical semantics in particular, that provide a basis for the evaluation process. Thus these chapters ground the evidential analysis in a larger context of text analysis.

This chapter focuses on the impact of the newspaper style on the interpretation of newspaper texts. Chapters 1 and 3 claimed that the complexity of the analysis of reported speech is significantly reduced when considering only the use of reported speech in newspaper articles, where it serves a specific purpose. This function of reported speech

in newspaper articles was identified as providing *evidence* for the embedded statement by citing the source. This chapter motivates these claims by grounding them in the newspaper style. While a comprehensive analysis of newspaper style is outside the scope of this thesis, particular features of newspaper style provide important information for the evaluation of reported speech.

Section 4.2 illustrates some linguistic conventions fro newspaper style, in particular conventionalized text structure and phrasal patterns. These features of newspaper style are well known and discussed in the literature.

Sections 4.3 and 4.4 ground the evidential analysis presented in the previous chapter in newspaper style. Section 4.3 reiterates the notion of two levels of information, the primary and the circumstantial information, which gave rise to the evidential analysis. Section 4.4 discusses the major linguistic and philosophical issues for the analysis of reported speech in general that do not arise in newspaper texts.

Section 4.5 finally introduces the notion of *stylized discourse* and its impact on the eventual interpretation of the reported speech. The analysis of texts with an underlying stylized discourse structure gives rise to an important distinction between lexical semantics and lexically triggered world knowledge.

4.1 Basic Assumptions

Newspaper articles have different styles and goals. Some differences depend on the style and audience of the newspaper itself, some depend on the function of the article in the newspaper (editorial, comment, newsbrief, newsstory, ...). To speak of *the* style of newspaper articles is therefore an oversimplification. Yet there are commonalities. One such common assumption is that newspaper articles serve to *inform* the reader on certain topics (unless marked otherwise). This information can be of a *fact*, an *event*, an *opinion*, or an *analysis*, among others.

Most newspapers are commercial products and are run under economical aspects. Additionally, some newspapers subscribe to a particular cause or affiliation. Chapter 3, page 43, pointed out that recognizing the resulting bias of the information selected and its presentation is an important part of understanding the newsstory. The assumed bias

of the newspaper has to be part of the reader's world knowledge or belief set. While no understanding of newspaper articles can be complete without this level of analysis, there remains still the underlying (simplified) assumption that newspaper articles report as objectively as possible. This assumption is as weak as the assumption underlying the Gricean maxim "Be informative." governing all communicative acts. But just as that maxim, the simplifying assumption that a newspaper article contains verifyable information is a very powerful tool for the analysis of news texts on purely linguistic terms. I make this assumption throughout the rest of this theses as a default that can be overridden by additional information available to the interpretation program.

4.2 Linguistic Conventions

We recognize newspaper style, even out of the context of a newspaper, immediately and without hesitation. Yet the syntactic structures used are so varied that even part of speech determination presents a real problem for automatic analysis. The distinct style of newspaper articles is therefore not due to a simplified set of syntactic constructions. It is in fact so characteristic because it employs complex constructions, such as embedded sentences (among them reported speech) and heavy noun phrases, disproportionately more often than is usual in spoken language or fiction. To a human reader this extra level of complexity is, however, not an obstacle but often a useful tool for efficient knowledge extraction. Heavy noun phrases, for instance, provide a compact characterization of a discourse entity, usually at the beginning of a sentence, *at the level relevant for the article.*

Conventionalized patterns, no matter how complex, help the reader to navigate through the information presented. Newspaper articles exhibit highly conventionalized patterns at all levels.

4.2.1 Conventionalized Text Structure

It is well known that an article consists of a *headline* and the *body.* Within the body we can distinguish the first sentence, which usually summarizes the major point of the article or sets up the major conflict addressed in the article (see [Lundquist, 1989] for

an elaboration of the latter point). The rest of the body elaborates the topic set up by the headline and the first sentence, putting the most salient information first, so that editing can be done with a pair of scissors in a pinch.

It is not possible to characterize newspaper style comprehensively in this thesis. Literature in text analysis has outlined aspects of newspaper texts, most notably [van Dijk and Kintsch, 1983] outline *superstructure* and *macropropositions* for news articles and exemplify their analysis on a story from Newsweek.

Superstructures are defined by van Dijk and Kintsch as schematic or conventionalized structures:

> "Typical of all these structures is their schematic nature: They consist of conventional categories, often hierarchically organized, that assign further structure to the various levels of discourse. Sometimes the categories will only affect surface structures, as in metrical or prosodic patterns, but semantic or even pragmatic units are often schematically organized as well. As in all these cases the structures seem to go beyond the usual linguistic or grammatical organization of discourse, that is, to be somehow additional or grafted onto the linguistic structures, we call them *superstrauctures*." [van Dijk and Kintsch, 1983, p. 236]

For the schematic strategies of assigning superstructures, the authors consider cultural information, social context and interaction, and pragmatic information. Superstructures are inferred from the *semantic macrostructure* embodied in *macropropositions*. Some example macropropositions for their example Newsweek text are: "There are no political choices", corresponding to the superstructure category "headline" (text: "Guatemala: No Choices"). Macroproposition: "The political situation in G. is more extreme than in ES[El Salvador]", superstructure category: "Lead".

These notions parallel the intuitions outlined in this thesis, especially the *style sheet* introduced in Chapter 7. The approach and level of description, however, are at the two opposing ends of the spectrum: van Dijk and Kintsch approach the text analysis top down, developing general high level strategies. This thesis outlines a bottom up strategy that projects up from the lexical semantic level. Focusing on the contribution of lexical semantics in the overall text structuring process also leads to a slightly different categorization of the kind of knowledge that van Dijk and Kintsch described in their superstructures. Section 4.5 and the following chapters show that much of this conventialized knowledge is implicit in the lexical semantics and can therefore not be relegated

outside the linguistic analysis. Thus while the analysis of van Dijk and Kintsch[1] touch on many related issues, it cannot be exploited directly here.

4.2.2 Phrasal Patterns

Certain repetitive information, such as weather forecasts and stock market activities, is reported in such highly constrained patterns that [Smadja and McKeown, 1990] have developed a special tool to extract the *phrasal patterns* that make up these reports. Information of this kind is sometimes presented graphically; if it is presented in "prose", the employed phrasal patterns are of a very small number and form, essentially forming a template.

It is important to recognize frequent phrasal patterns in these contexts for two reasons. First, recognizing a phrasal pattern and assigning a predetermined meaning to the entire pattern is faster and easier than applying compositional semantics. But secondly, the meaning of the pattern in its context may be much more specific than general compositional semantics indicates, in fact these patterns constitute part of the *sublanguage* for that genre of articles.

Reported speech forms a much looser pattern that, nontheless, has a particular meaning in newspaper articles.

4.3 Two Levels of Information

Chapter 1 suggested that reported speech is so very frequent in newspaper articles because the American newspaper tradition requires sources to be identified (or have to be identifyable in liability suits), information to be verified, and the facts to be presented in an objective (here contrasting with suggestive) way. This tradition has even been transferred to television news shows, where the anchor reads news that can be objectively verified, solicits further (in-depth) information from the individual reporters on location, who summarize the circumstances and provide opinions and first hand information in

[1]The same holds for other research in text analysis, which has greatly inspired my work but is not directly applicable.

form of interviews with people pertaining to the story. It is interesting to note that even when the journalist on location has experienced the event personally, he or she will not report first hand but will usually be interviewed by another journalist *as if they were like any other eyewitness*, not a journalist who could very well do the whole story by himself or by herself. This indicates that in American journalistic practice it is not felt proper to conflate the two roles of reporter and witness or source of information.

One reason why the conflation of one person as reporter and witness might cause confusion in television (or radio) news is that two levels of information have to be distinguished, namely the *primary information*, usually provided by an expert or witness interviewed, and the *circumstantial information*, provided by the journalist.

In newspaper articles these two levels are distinguished by different means. In the case of reported speech, the distinction is one of syntax. We find the circumstantial information in the matrix clause and the primary information in the complement clause.

Primary information is that information which pushes the newsstory forward. *Circumstantial information* is meta-information, which embeds the primary information within a perspective, a belief context, or a modality. For tasks such as knowledge extraction it is the primary information that is of interest. For example in the text of Figure 4.1 (repeated from Chapter 3) the matrix clauses (italicized) give the circumstantial information of the *who*, *when* and *how* of the reporting event, while *what* is reported (the primary information) is given in the complements.

US Advising Third Parties on Hostages

(R1) *The Bush administration continued to insist yesterday that* (C1) it is not involved in negotiations over the Western hostages in Lebanon, (R2) *but acknowledged that* (C2) US officials have provided advice to and have been kept informed by "people at all levels" who are holding such talks.

(C3) "There's a lot happening, and I don't want to be discouraging," (R3) *Marlin Fitzwater, the president's spokesman, told reporters.* (R4) *But Fitzwater stressed that* (C4) he was not trying to fuel speculation about any impending release, (R5) *and said* (C5) there was "no reason to believe" the situation had changed.

(A1) Nevertheless, it appears that it has. ...

Figure 4.1: Boston Globe, March 6, 1990

The particular reporting verb also adds important information about the manner of the original utterance, the preciseness of the quote, the temporal relationship between matrix clause and complement, and more. In addition, the source of the original information provides information about the reliability or credibility of the primary information.

4.4 Conventionalized Pragmatic Constraints

Newspaper articles feature a very limited subset of language's potential; in particular the pragmatic aspects of speech acts, intensionality, etc. are by convention very restricted indeed. This section will illustrate some of the conventionalized restrictions without attempting to give a detailed analysis of newspaper style.

4.4.1 Speech Acts

There is only one permissible speech act for articles[2], namely *inform*. There are two different levels of information, namely the "objective" information presented in normal articles and the subjective information or commentary presented in the editorials. Editorials have their very own style, which does not concern us here.

To say that a whole newspaper embodies one speech act, namely *to inform*, is an interesting Gedankenexperiment, leading us to consider the newspaper as a whole in the context of previous issues. This global view is common in political analysis and beyond the linguistic realm. The function then of all the parts of a newspaper is similarly to inform. This fact is normally taken for granted, made explicit only by pragmatic theories, such as Grice's maxim [Grice, 1967] *be informative*, which is supposed to underly all cooperative communication. The speech act quality of reported speech is however of special importance for the correct interpretation of the primary information,

[2]Announcements, requests, etc. which one might find in a newspaper are not considered here for two reasons, (a) because they do not occur within an article (and usually occur in sections that are clearly marked for that purpose, such a obituaries, apartment rentals, etc.) and (b) because these are only secondary announcements, requests, etc.; the newspaper does not announce or request, it *informs* its readers about the announcement, request etc. that somebody else makes.

because the implicit *commitment* of the newspaper to provide *correct* information. The speech act quality is also present in all casual uses of reported speech, where the fact that an utterance has been attributed to a source binds the reporter to a faithful and correct paraphrase, if not a literal[3] rendition, of the reported material *along with the necessary contextual information to put the utterance into perspective*. The speech act component of reported speech is therefore part of the semantic class presented in Chapter 7.

On a different level reporting verbs can of course lexically denote speech acts that are then attributed to the source, examples are *announce, agree, pledge, etc.* This encoded speech act is part of the decompositional meaning of the individual reporting verbs. The only special inferences that can be drawn from these speech act encoding reporting verbs is that *announce* usually refers to an event in one of a few settings, including a press conference or a press announcement. This is heuristic world knowledge and will not be further considered. Thus a full blown speech act analysis is not required for the basic text representation of newspaper articles (as opposed to a complete interpretation). The default speech act *inform* for reported speech will be incorporated in the meta-lexical structure described for reporting verbs in Chapter 7.

4.4.2 De Re and De Dicto Reference

Chapter 1 already introduced the problems of de re and de dicto reference and that they are not of importance for newspaper articles (cf. also [Bach and Harnish, 1979, p. 30]). Let me illustrate this claim with some data from the Wall Street Journal corpus:

(1) (a) The Toronto-based real estate concern said each bond warrant entitles the holder to buy C$1,000 principal amount of debentures at par plus accrued interest to the date of purchase.

(b) SHAREDATA Inc. said it will amend a registration statement filed with the Securities and Exchange Commission to delete a plan to sell 500,000 newly issued common shares.

(c) In his first state of the nation address, Salinas pledged to continue his program of modernization and warned opposition politicians that impeding progress could cost them popular support.

(d) Sen. Kennedy said in a separate statement that he supports legislation to give the president line-item veto power, but that it would be a “reck-

[3]The meaning of a quote in newspaper texts is no longer a literal rendition, but a paraphrase of what the source might have said. I will ignore this complication here.

less course of action" for President Bush to claim the authority without congressional approval.

(e) John Bolton, the assistant secretary of state for international organizations, told Congress that the continuing "statist, restrictive, nondemocratic" programs make rejoining any time soon "extremely unlikely."

(f) In a statement, Jaguar's board said they "were not consulted about the (Ridley decision) in advance and were surprised at the action taken."

(1a) is a typical case of Wall Street Journal text, where all references are without a doubt de re. (1b), too, is of a frequent pattern in the corpus, namely $<$ *company* $>$ *said it* $<$ *action* $>$, where the impersonal pronoun refers back to the company in the subject NP of the matrix clause. (1c) is a case where the NP *his program of modernization* could conceivably be construed as de dicto, assuming an opposition that disagrees with the program in fact implementing *modernization*. But this reading is very far fetched, because the NP *his program of modernization* is a name that refers to a program that can (de re) be identified based on that name, disregarding the name's descriptive adequateness. The way that de dicto references are rendered is exemplified in (1d), (e), and (f). The material given de dicto is enclosed in quotation marks. Notice, however, that this literally quoted material is not of the same quality that concerns philosophers when they discuss de dicto references: the references are not factually false or hard to decipher outside their original utterance context. The material enclosed in quotation marks rather contains a value judgement which the journalist does not want to endorse or it contains the literal phrasing of a comment to avoid misunderstanding. De dicto references are not a problem in newspaper style.

4.4.3 Reader's Belief about Author's Intention

This problem, usually considered central in speech act and belief literature, is also not of primary importance for newspaper articles. The assumption is that a reader reads a newspaper to get information; that the reader expects the newspaper to provide information; that the reader chooses a particular newspaper because he or she can rely on the provided information. By default the reader does not challenge the assumption that the intention of the author is to inform the reader about important events. The fact that readers are likely to dismiss entirely what is written in a newspaper that is not in agreement with the reader's own point of view only reinforces the default case.

If the intention of the author is deemed to be relevant, it can be recovered from both the circumstantial and the primary information. Circumstantial information will indicate which side the author favors or disfavors. Primary information, compared with information gleaned from different sources, will reveal the truthfulness and completeness of the information. Again, as a default an author cannot violate the general point of view of the newspaper overall, unless he or she publishes it as an editorial[4].

Much more relevant than the intention of the author is the point of view and the motivation of the reader. As Hobbs observes:

> . . . not only is the role of the author's or speaker's intention indirect; it is frequently not very important.
>
> The agent's interpretation procedure works by drawing inferences from its belief system . . . [Hobbs, 1990, p. 19]

The model of interpreting newspaper articles is much closer to information retrieval and incorporation, than to cooperative discourse, and I will in the following assume that the goal of the (human or automated) reader is to retrieve *a certain kind of information for a special purpose*. An example is the collecting and representation of information about mergers and acquisitions in a particular format.

I will come back to the point of analyzing texts from different perspectives in Chapter 8, where a general text representation scheme will be presented that can be used by several different "user programs" that will *interpret* the text. As a simple model consider [Hobbs, 1990], that says that the interpretation process F of a text T under the set of beliefs K yields a representation I of the text, expressed formally:

$$\mathrm{F(K,T)} = \mathrm{I}$$

Hobbs points out that for interpreting literary texts both K and I are unknown, yet constrain each other. For the interpretation of newspaper texts, K is much more constrained than for general fiction. I follow Hobbs and assume that knowledge about language and the interpretation of texts of different styles and periods resides in K. K

[4]This whole discussion is of course moot for letters to the editor: the letter writer by assumption does not have any obligation to inform or adhere to the paper's point of view, in fact, letters to the editor often take issue with the paper's view on a given topic.

contains two distinct sections, namely the idiosyncratic section (which gives a reader his or her particular point of view) and the socially normed section, where we find conventionally determined knowledge such as lexical meaning, semantic procedures, and default knowledge about newspaper style. The idiosyncratic section contains true "beliefs" and memory of events that center around the individual. I will here only be concerned with the socially normed section and within this section only with the knowledge associated with language. One form of conventionally normed knowledge is knowledge about stylized discourse structures.

4.5 Stylized Discourse Structure

Newspaper articles are not only often based on some form of prior discourse, but usually *reflect* the structure of the underlying discourse implicitly. Of special importance is a type of discourse situation which I will call *stylized discourse*, a form that is conventionalized to very rigid forms, such as interviews, court and Congressional hearings, press conferences, etc. Stylized discourse is very rule-governed, the topic is known and well defined, turn taking is usually signalled, the participants fill certain roles which in part define their underlying intentions. The analysis of newspaper articles which report on stylized discourse situations often requires a rough reconstruction of the underlying discourse situation.

Senate hearings, press conferences, court proceedings, scientific conferences, panel discussions — all these forms of dialog are highly constrained by conventions; some have their own sublanguage, e.g. "legalese"; none conform to normal rules of cooperative discourse ("Be cooperative." "Be relevant'." "Be brief." See [Grice, 1967].) This exemption from normal rules of cooperative behavior is important to recognize, not only when observing (or being involved in) a stylized discourse situation, but also when reading an article reporting on such a stylized discourse. In analogy to task-oriented forms of discourse [Grosz and Sidner, 1986], I characterize stylized discourse as *style-oriented*.

Participants in stylized discourse play certain *roles* (e.g. lawyer, defendant, financial expert, interviewer), their interaction is largely predetermined, often mediated by a *moderator* who announces whose turn it is, what the role of the speaker is, and what

legitimation the speaker has to speak on the topic. There is usually only one single topic; if there are more topics, they are carefully introduced. Usually stylized discourse serves to *inform an audience from different perspectives.*

One difficulty for the interpretation of stylized discourse is the fact that the personal intentions of the discourse participants may be at odds. In an interview situation, for instance, it may be the interviewer's goal to get the interviewed person to talk about a third person, about whom the interviewed does not intend to disclose any information. In some forms of stylized discourse these conflicts of interest are predictable. For example, court proceedings have as basic roles two conflicting parties; one, being charged by the other of a crime or wrongdoing, trying to defend itself, the other trying to support the charge.

It is this possible conflict of interest that requires one not to take the individual statements at face value, but to *interpret* and *evaluate* them with respect to the *role* and private *intentions* of the speaker. The importance of assessing the *reliability* of the speaker's utterance through the hearer has been discussed in many different contexts. Most relevantly [Cohen, 1987], when discussing evidence determination in argumentative discourse, mentions that differences in beliefs between hearer and speaker can force the hearer to adopt a *"hypothetical person's" belief* where the speaker's utterance only makes sense in light of a (set of) belief(s) that the hearer does not (necessarily) share.

The assessment of the underlying beliefs and intentions is based on world knowledge in its most general form. The assessment of predictable role behavior in a stylized discourse situation, on the other hand, is constrained by conventions. These conventions form a kind of world knowledge that can be aptly represented in *scripts* [Schank and Abelson, 1977]. Stylized discourse scripts are different from general world knowledge because an understanding of these scripts underlies the the very definition of the words denoting particular roles in the stylized discourse situations (*plaintiff*, *defendant*). The close connection between lexical semantics and world knowledge has of course not only been observed by Schank, but has also led [Hobbs *et al.*, 1987] to claim that a distinction between lexical semantics and world knowledge is not possible.[5] I argue, in contrast, that the distinction is necessary for a robust analysis of large amounts of text, where

[5]Stylized discourse scripts are of course also related to van Dijk and Kintsch's superstructures. The fact that research from such different areas has addressed this problem stresses its importance.

a full commonsensical model of all relationships expressed is not (in the near future) feasible.

Let me illustrate the problem with an example. Figure 4.2 shows a short text from the Wall Street Journal corpus concerning a court decision.

> THE CASE OF THE FAKE DALIS: In federal court in Manhattan, three defendants pleaded guilty to charges of fraud in connection with the sale of fake Salvador Dali lithographs. James Burke and Larry Evans, formerly owners of the now-defunct Barclay Gallery, and Prudence Clark, a Barclay sales representative, were charged with conducting high-pressure telephone sales in which they misrepresented cheap copies of Dali artwork as signed, limited-edition lithographs. The posters were sold for $1,300 to $6,000, although the government says they had a value of only $53 to $200 apiece. Henry Pitman, the assistant U.S. attorney handling the case, said about 1,000 customers were defrauded and that Barclay's total proceeds from the sales were $3.4 million. Attorneys for Messrs. and Evans and Ms. Clarke said that although their clients admitted to making some misrepresentations in the sales, they had believed that the works were authorized by Mr. Dali, who died in January. The posters were printed on paper pre-signed by Mr. Dali, the attorneys said.

Figure 4.2: Underlying "legal court" discourse situation.

This article is not structured in an obvious fashion. It is not chronologically structured, it does not develop a topic, it jumps from one source to another. Yet it makes sense, thanks to our understanding of the underlying situation.

A much simplified script for the "legal court" situation is given in Figure 4.3[6].

In this ordered fashion it is clear what the underlying structure of the article itself was: first summarizing the decision the journalist went on to introduce the charge, using reported speech to introduce speculative material by the U.S. attorney, and finally presented the "defense" line through a quote from the defendant's attorney.

Also clear is the evaluative context of the different instances of reported speech. Both attorneys attempt to portray an extreme of their client's positions. A pre-signed poster (even without any "art" printed on it) is worth more than $53 after the artist's death. And no good art dealer, who truly believes to have authorized signed lithographs would sell them off hastily soon after the artist's death, when it is expected that authorized

[6]Missing from the presented script fragment is of course all temporal information about the sequence of events, association between parties and events, and subevents.

Court Proceedings:

Participants:

Defendant Side:

Defendant: James Burke, Larry Evans, Prudence Clark

Counsel to the defendant:

Plaintiff side:

Plaintiff: U.S.

Counsel to the Plaintiff: Henry Pitman

Court:

Judge:

Jury:

Events:

Charge: Conducting high-pressure telephone sales in which they misrepresented cheap copies of Dali artwork as signed, limited-edition lithographs. The posters were sold for $1,300 to $6,000, although the government says they had a value of only $53 to $200 apiece. About 1,000 customers were defrauded and that Barclay's total proceeds from the sales were $3.4 million.

Defense: Attorneys for Messrs. and Evans and Ms. Clarke said that although their clients admitted to making some misrepresentations in the sales, they had believed that the works were authorized by Mr. Dali, who died in January. The posters were printed on paper pre-signed by Mr. Dali, the attorneys said. Pleading guilty.

Ruling: Guilty of fraud in connection with the sale of fake Salvador Dali lithographs.

Figure 4.3: "Legal court" script fragment.

lithographs will rise in value over time. Thus both attorneys' statements do not make sense taken at face value. They have to be interpreted within the framework of the legal argument, where both parties are expected to state extreme positions.

Another example for the importance of dectecting the underlying stylized discourse situation is illustrated by the text in Figure 4.4, where we are presented with one side of a reportedly "heated" dialog between President Bush and a reporter at a news conference. From Bush's answers and some background knowledge we can reconstruct what the questions were about, thus enabling us to *evaluate* the answers in the right context.

An emphatic Bush again bars rebel aid

Associated Press. NEWPORT BEACH, Calif.- President Bush yesterday emphatically ruled out US military help for the rebels fighting Saddam Hussein, saying he will not put "precious American lives into this battle."

In a heated reply to a question, Bush said, "I made very clear from day one it was not the US objective to get Saddam Hussein out of there by force." He said "I have not mislead anyone" about US policy in Iraq.

He expressed revulsion for Saddam Hussein's attacks against Kurdish and Shiite opponents, but said they never expected US support for their efforts to oust him.

Speaking at a news conference, Bush said, "We fulfilled our obligations" under United Nations resolutions to remove Iraq from Kuwait.

...

Figure 4.4: Stylized discourse situation "news conference".

In this text it is important to evaluate Bush's statements in light of a news conference, where we expect pointed questions and where Bush being "emphatic" would not mean a diplomatic crisis. The first two paragraphs would receive a very different evaluation if they had been reported happening at a U.N. gathering or at a White House dinner party.

The two texts point out very clearly that world knowledge is necessary to evaluate reported speech. This world knowledge must contain knowledge of stylized discourse situations. It is outside the scope of this thesis to define stylized discourse situations. Yet from the discussion above it is possible to suggest that some basic script-like knowledge of the structure of stylized discourse be linked with the lexical entries that are defined only within these situations.

This structure overlaying the lexical entries can also help to define the (opposing) supporting groups defined in Chapter 8. However, as I discuss in Chapter 9, basic redundancy in newspaper articles often allows the linguistic analysis to detect the (opposing) supporting group structure even when no general stylized discourse information is available.

Chapter 5

A Computational Lexical Semantics

The previous chapter discussed how the newspaper style constrains certain pragmatic issues which are important for the final evaluation of an evidential analysis. This chapter begins a discussion of the lexical semantics of the most important parts of the evaluative environment, namely the reporting verb and the source description. This chapter introduces the framework of the Generative Lexicon, in which the lexical semantics is expressed. The methodology of deriving the lexical semantics is guided by data from corpus analyses (more on corpus analysis techniques follows in the next chapter). The chapter concludes with an example that illustrates both the power of the generative lexicon and of a usage based lexical semantics of reporting verbs.

Traditional semantics assumes that the lexicon provides the information necessary for a particular analysis. We can in fact view most semantic theories as characterizing the properties of a word or a class of words the lexicon must contain for a particular theory. Unfortunately, these "lexical semantic features" of words under different theories are not cumulative; often they might even be contradictory. Recently, however, there have been efforts to build lexicons that are *consistent, comprehensive,* and *theory neutral.* This change has been made possible in the computational linguistics community by the accessability of on-line, machine readable dictionaries and large, on-line text corpora from different sources. These corpora provide the tools to build *general purpose, reusable*

lexicons based on usage in real text (cf. [Boguraev and Briscoe, 1989]).

The availablity of more diversified sources on-line also enables computational linguists to construct coherent sublanguage dictionaries for very restricted domains that were previously not thought to be useful, such as the style and language of one particular magazine or the use of metonymy in the computer science community. This is in distinction to the efforts of traditional corpus linguistics, where the goal is to construct a *balanced* corpus that contains "a little bit of everything". Given the small size even of the "large" Brown corpus [Francis and Kučera, 1982] of one million words, the balanced corpora were not able to provide sufficient data for all the (contextually) different meanings of a word. The discussion of some data in Section 6.4.3 suggests that a narrow (single source) corpus might provide deeper insights into the behavior of words in a given *style* or *context*. The behavior of the same word in different contexts can then be studied, allowing generalizations *and* contextual particularizations to be formulated in one concise entry. For a preliminary suggestion of the feasibility of such sublanguage acquisition, see [Hindle, 1990, Anick and Pustejovsky, 1990]. For a discussion of the general problems of sublanguage acquisition, see [Grishman *et al.*, 1986]

The lexical semantics developed in this thesis is cast in the formalism of the *Generative Lexicon Theory*, outlined briefly in the next section, followed by a discussion of the methodology used in the following chapters to derive the lexical semantics for individual words based on the analysis of two large on-line corpora.

5.1 Generative Lexicon Theory

The *Generative Lexicon* is a formalism currently under development at Brandeis University based on [Pustejovsky, 1991]. The goal is to provide a representation language that can adequately describe a lexical semantics that not only defines the meaning(s) of a word but also provides a structure that allows for dynamic construction of clusters of related words — antonyms, synonyms, collocational information and more. The Generative Lexicon (GL) is an interface between linguistic (i.e. syntactic and semantic) terminology and commonsense and real world concepts, enabling independent yet integrated processing of real texts on these different levels. In the philosophy of GL a richer lexical semantics yields a more compositional semantics that can project from

the lexical level upwards, even onto the level of *text*.

5.1.1 Vocabulary of GL

GL is both a theory of lexical semantics and a knowledge representation formalism. The motivating idea behind GL is that traditional static lexicons are not adequate for computational use and that a *dynamic* representation formalism allows one to interconnect entries for different words, thus creating a lexical network of entries, rather than a list[1]. To illustrate the concerns of GL, consider the case of polysemous words. Traditional lexicons list a definition per word meaning, for instance *bank* will have an entry referring to (1) *a piled up mass (as of cloud or earth) ...* or (2) *an establishment concerned esp. with the custody, loan, exchange, or issue of money, ...* [Woolf, 1974]. The case is clearcut in cases of homonyms, as for these two meanings for *bank*. But for polysemous words whose meanings are *semantically* related, traditional lexicons list a selection of the more frequent contexts and idioms in which the polysemous word can occur as different subentries (often illustrated with an example sentence). Depending on size, audience, and technical bias, however, this selection appears arbitrary and in fact is often not sufficiently motivated for non-native users of dictionaries as well as for non-intended user programs of computational lexicons. GL entries have a richer structure into which the word meaning is decomposed, in fact implying by their very structure several "polysemous" uses of the word, as for instance the use of *bank* (monetary) in the following examples (cf. [Pustejovsky, 1995]):

(1) (a) The bank raised its interest rates yesterday. (i.e. the institution)
(b) The store is next to the new bank. (i.e. the building)

GL is able to represent the lexical entry for *bank* distributively in different *qualia* roles, indicating that a *bank* is an *institution* that is housed in at least one *building*, consisting of *management* and *customer services*, where the management (i.e. *president, vice-president, board of directors, shareholders, ...*) decides on *financial policy* and the customer service department (i.e. *bank tellers, ATMs*) performs the financial services. Moreover, GL encourages to slim down the lexical entries by refering to superconcepts

[1]The idea of interconnections between words is of course not new in the AI literature; for an early suggestion see [Quillian, 1968].

from which information can be inherited. For the example for *bank* we can assume that all customer service institutions will share much of the structure of the outlined entry.

GL provides many different devices for the representation of a word entry. Every lexical entry carries information relating to four different aspects of its meaning, namely:

1. **Argument Structure:** The behavior of a word as a function, with its arity specified. This is the predicate argument structure for a word, which indicates how it maps to syntactic expressions. (see [Williams, 1981, Grimshaw, 1990])
2. **Event Structure:** Identification of the particular event type (in the sense of [Vendler, 1967]) for a word or phrase: e.g. as state, process, or transition.
3. **Qualia Structure:** The essential attributes of an object as defined by the lexical item.
4. **Inheritance Structure:** How the word is globally related to other concepts in the lexicon. This includes traditional, *fixed* inheritance (cf. [Touretzky, 1986]) and *projective* inheritance (cf. [Pustejovsky, 1991]).

According to [Pustejovsky, 1991], *"These four structures essentially constitute the different levels of semantic expressiveness and representation that are needed for a computational theory of lexical semantics. Each level contributes a different kind of information to the meaning of a word. The important difference between this highly configurational approach to lexical semantics and feature-based approaches is that the recursive calculus defined for word meaning here also provides the foundation for a fully compositional semantics for natural language and its interpretation into a knowledge representation model."*

It is the *qualia structure* that captures the decompositional meaning of a word and I will only refer to the qualia structure of lexical entries in this dissertation. The qualia structure is a system of relations that characterizes the semantics of nominals, very much like the argument structure of a verb. It is based on Aristotle's theory of explanation and ideas from [Moravcsik, 1975] and consists of four *qualia roles*:

Constitutive Role: the relation between an object and its constituents, or proper parts, i.e. *material, weight, parts and component elements.*

Formal Role: That which distinguishes the object within a larger domain, i.e. *orientation, magnitude, shape, dimensionality, color, position.*

Telic Role: Purpose and function of the object, i.e. *the purpose that an agent has in performing an act or the built-in function or aim which specifies certain activities.*

Agentive Role: Factors involved in the origin or "bringing about" of an object, i.e. *creator, artifact, natural kind, causal chain.*

I will not be using the full theory behind GL in this dissertation; I will use the qualia structure informally, details should become clear with the examples. For more detail see [Pustejovsky, 1991, Pustejovsky, 1995].

5.2 A Methodology for Corpus Based Lexical Semantics

The methodology presented here for deriving lexical structures is a combination of corpus analysis and linguistic theory. Hypotheses are confirmed in part with the analysis of large on-line corpora. Correlating data from semantically close words from the corpus also gives rise to linguistic hypotheses. This two-way relation of inspiration between linguistic intuitions and linguistic theory on the one hand and the data from corpus analysis on the other proves very fertile, yet is very hard to formalize. I will therefore not propose mechanisms to automatically extract components of the lexical semantics of lexical entries, but will focus on those aspects of some reporting verbs and source descriptors that are interesting *because* they cannot be entirely extracted automatically but will also not be apparent when considering the corresponding words in isolation. The key idea here is that paradigmatic behavior of a word can only be detected when compared to the behavior of other words in the same semantic field [Trier, 1931] (see Section 7.2 for more detail). [Bergler, 1991] reported that interesting aspects of a set of seven reporting verbs emerge when comparing their relative preference for different *degrees of animacy* over their subjects. Interestingly, the patterns for semantically close verbs were more similar than for semantically more distant words.

Thus, intuitive notions about a semantic field combined with actual data of the usage of the words under consideration gives rise to initial linguistic hypotheses, as illustrated in Figure 5.1.

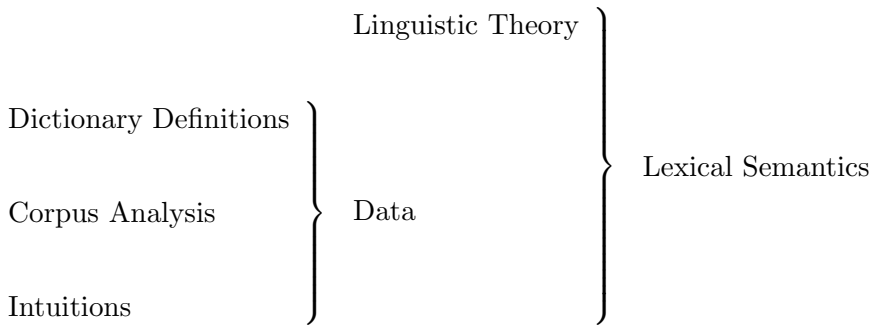


Figure 5.1: Methodology for a corpus-based lexical semantics

On the other hand, testing linguistic hypotheses on the data derived from corpora can yield a refinement of the linguistic intuitions. Section 6.4.1 discusses how the notion of *discourse polarity items* emerges from data testing the cooccurrence of negative markers with *insist*.

5.2.1 Lexical Semantics of a Semantic Field

To study different words of a semantic field together allows us to perceive 'deep' semantic properties and distinguish them from accidental meaning components of individual words. Both are important for the understanding of a word; their distinction is important when judging semantic closeness of 'synonyms' and preciseness of antonyms in a given context. However, the study of a semantic field also reveals whether certain syntactic constructs are more closely connected to a semantic concept or whether there are any collocational restrictions on arguments, etc., beyond the selectional restrictions.

A lexical semantics for a semantic field must, therefore, address the following points:

- Define the semantic field, i.e. the semantic commonality that all members of the field share[2].
- Define syntactic and other structural commonalities *typically* shared by members of the field.

[2]This is similar to the notion of *prototype* (cf. [Rosch *et al.*, 1975]).

- Define the structure of the semantic field, i.e. give the relationships of the members towards other members ([Véronis and Ide, 1991]).
- Define a set of semantic dimensions along which the individual words are distinguished.
- Define the individual words *as they deviate from the general pattern.*[3]

This approach requires the lexicon to have a well-structured *meta-lexical* level, on which generalizations over groups of entries can be defined. It also requires an inheritance structure to pass down general information. GL provides both. I will here not discuss inheritance structures but will assume mechanisms as described in [Pustejovsky, 1991].

The meta-lexical structure can in part be mapped onto the GL construct called *Lexical Conceptual Paradigm* (LCP). LCPs provide the interface between linguistic notions and conceptual notions. [Anick and Pustejovsky, 1990] introduces LCPs that capture different syntactic patterns. The advantage of LCPs over traditional mappings from syntax to the word entry lies in the fact that LCPs are somewhat dissociated from the individual entry, allowing the same LCP to serve for all entries that pattern accordingly. Thus LCPs naturally implement a meta-lexical level as needed for the description of a semantic field. Meta-lexical structures for the semantic field of reporting verbs will be presented in Chapter 7.

5.3 Cocompositionality

Generative Lexicon theory incorporates a richer lexical semantics than traditional approaches. Of particular importance is a new appreciation of the role of word classes in the compositional process.

Traditional semantic theories share the basic control structure of the semantic composition process for simple sentences. The verb assumes the active role of organizing

[3]The observation that words in a field are more crisply defined *negatively*, i.e. in terms of their deviation from the general pattern than *positively*, i.e. listing their individual properties can be found in [Trier, 1931].

noun phrases and prepositional phrases according to its argument structure, thereby *assigning* the meaning to these arguments (such as case roles [Fillmore, 1968]). A recent example for this view is [Grimshaw, 1990]. Assuming this rigid control structure requires a very strong and powerful level of pragmatic or commonsense recovery for ambiguous cases or apparent violations of selectional restrictions. Consider the case of *logical metonymy* as defined in [Pustejovsky, 1991], where a subpart or related part stands for the word itself: *The newspaper announced a hiring freeze.* This obvious violation of the selectional restriction [+animate] for the subject of *announce* disrupts the normal composition process, flagging an error and requires a follow-up process that infers the intended metonymy.

Generative Lexicon theory breaks with this asymmetrical control structure and allows nouns to actively participate in the semantic composition process in the form of *cocomposition* [Pustejovsky, 1991]. The advantage of a more active role of nouns in particular holds not only for relational nouns (*operation*), but also for cases that have lead to a proliferation of different word senses for verbs because the type of the verb seems to be different for different argument types. Pustejovsky illustrates this with the example *bake a potato* vs. *bake a cake*, where *bake* has a change/state reading for *potato* but a create reading for *cake*. Pustejovsky's suggestion is to assume only one word meaning for *bake*, the process reading. To account for the create sense in the second example, the argument *cake* cospecifies the create reading, actively changing the process reading by superimposing its own definition as an artifact that comes into being by baking.

The principle of cocompositionality reduces the number of possible ambiguities that the semantic composition phase has to hand on to be resolved by a second stage, thus effectively reducing the need for one kind of commonsense inference.

The next section will elaborate on the benefits of cocompositionality on the example of reported speech. Lexical entries that reflect usage in newspaper articles combine to actively describe coherence constraints between the source and the complement for particular reporting verbs. These coherence constraints serve to focus commonsense inferences.

5.4 Coherence Constraints

Assuming the principle of relevance for the evidential use of reported speech means that (a) the content of the primary information has to be relevant within the topic being discussed and (b) the source must have some legitimation or relevance with respect to the topic. This is a basic coherence constraint and precludes quoting a former actor on matters of foreign policy unless some additional qualifying information is known, such as that actor being President at the time of utterance. This is in fact the role of the source descriptions in newspaper articles. The reporter indicates the relevance of the source to the topic in the lexicalization of the source description. Some reporting verbs, however, require additionally that stronger coherence relations hold between source and complement. One of these verbs is *announce*:

(2) a) The New York Stock Exchange announced the listing of two foreign banking firms.
b) ? Scientists announced the listing of two foreign banking firms on the stock market.
c) ? The New York Stock Exchange admitted the listing of two foreign banking firms.

Example (2b) is questionable because scientists are not qualified to announce for the stock market, whereas (c) is odd because *admit* in the sense of "no longer refuse to admit to" is not proper conduct for the NYSE, where all listings have to be made public immediately.

Let us consider coherence constraints between reporting verbs and their sources by first investigating its impact on the frequent use of metonymy[4].

As reported in [Bergler, 1991], reporting verbs frequently occur with a particular kind of metonymy, where the violated selectional constraint is that the *source* of the utterance, i.e. the agent, is not a single individual human being, but a *city* or *institution* or even a *building*, as illustrated below:

(3) a) Marlin Fitzwater said ...
b) Washington said ...

[4]For a more detailed discussion of metonymy see Section 6.4.2 on page 114.

c) The White House said ...
d) IBM announced ...

All reporting verbs accept metonymic extensions where the original human source is replaced by some *significant* superclass they fall into. In the case of reporting verbs of special importance is the *employer* (*IBM*) or the membership in a *group* (*the French*). But we find also the characterization of the source in terms of a *role* he or she played , as in *the witness*[5]. This general or schematic "knowledge" can best be stored in a concept "REPORTING–VERB", along the lines of Def.1 on page 42 — all the verbs defined with a reporting verb sense will then inherit the "permission" to use metonymic extensions to a superclass or a specific role played in an event.

But this is not a sufficient restriction on all possible metonymies. Consider the following scenario: a German makes a disparaging remark ϕ about the German government in the United States.

The types of expected descriptions of the source, that would be appropriate, are:

(4) a) Hans Glück said that ϕ (Name)

b) A German said that ϕ (Nationality)

c) ? A Mercedes Benz employee said that ϕ (Affiliation)

however, given what ϕ denotes, the following descriptions would be deemed inappropriate and in fact uninformative (cf. [Grice, 1967], [Gazdar, 1979])

(5) a) ? A human said that ϕ

b) ? A vegetarian said that ϕ

c) ? A European said that ϕ

This example illustrates that not all true descriptions of a source can be used as subject in reporting contexts. The source has to be *relevant*[6] to the topic of the com-

[5]Note that these frequent metonymic extensions are implicit in the descriptive semantic grammar in Section 3.3.1.

[6]I will not attempt to define relevance here. See [Sperber and Wilson, 1986] for extensive discussion of the notion of relevance and [Wilks, 1986] for a criticism of their approach.

plement clause. The lexical realization of the source in newspaper articles generally adds to the evaluation of the *credibility* or *reliability* the source has with respect to the topic. Thus a German in general has more competence and thus reliability on issues concerning the German government because of direct affectedness; this fact however is lost if we characterize him (correctly) as a European.

This is little surprising; general pragmatic considerations such as Gricean maxims ("Be relevant", "Be informative" [Grice, 1967]) account for this fact. It has been noted that it is extremely difficult to put these pragmatic considerations into computational form [Wilks, 1986]. We will show here that some pragmatic coherence can, however, be established on the basis of lexical semantics.

The dilemma here is that a basically pragmatic problem is part of the selectional restrictions[7] of reporting verbs and that therefore the occurring systematicity has to have some reflection in a lexical semantics. I will sketch informally how the coherence constraint works in the framework of GL.

Let us first state the problem with an example. *Announce* selects for a subject that has some "legitimization" for making the statement that is repeated or paraphrased in the complement[8]. The relation between subject and complement that constitutes the "legitimization" is established by the *projective conclusion space*, a device that generates certain semantic relationships between lexical items dynamically, much in the way a semantic net records semantic relationships between words statically.[9]

The lexical definition of *official* is in the American Heritage Dictionary is:

official adj

1. Of, pertaining to, or authorized by a proper authority; authoritative.
2. Formal or ceremonious: *an official banquet.*

n

1. One who holds an office or position.
2. A referee in a sport.

[7] I take here a different view from [Grishman *et al.*, 1986]. There cooccurrence between S–V–O was viewed as a syntactic selectional constraint in a subdomain that had to be fully listed. I assume that more general, metalevel rules can be described that allow to capture many of the same type of data.

[8] See Chapter 7 for a more detailed discussion of the lexical semantics of the words involved here.

[9] For detail, see [Pustejovsky, 1991].

and in the Longman Dictionary of Contemporary English we find:

> **official** **adj** of or about a position of trust, power, and responsability: *an official position/an official occasion/an official manner of speaking.* — opposite: **unofficial**; compare officious

which results in a GL entry of the form[10]:

official(*x*, y)
[Form: human(*x*), organization(y), position(z) & in(z,y) & hold(*x*, z)
& has-authority(z)]
[Telic: work-for(*x*, y), trust(y, *x*)]
[Constitutive: individual(*x*)]
[Agentive: inherit-from(y)]

announce(A,B)
[Form: REPORTING–VERB]
[Telic: assert(A, B) & new(B)]
[Agentive: legitimation(A, utter(B))]

I am ignoring the speech act reading of *announce* here entirely. *New(B)* refers to the semantic dimension considering the status of the proposition B in the text. *Announce* implies that the information is new.

The combination *Officials announced* sounds very appropriate, almost idiomatic, regardless of the content of the complement clause, because the semantic selectional restriction imposed by *announce* is met *at the lexical semantic level* by the definition of *official* as somebody of authority or trust, implicitly "legitimized". This is why *Officials announced* sounds more like a phrasal pattern than like an ad hoc generative combination.

This is an obvious case where a simple feature matching procedure can establish the coherence relation given two lexical entries. Consider the slightly more complex

[10] *Form, Telic, Constitutive,* and *Agentive* are described in Section 5.1.1. For the purposes of this discussion they can simply be glossed as: An *official*, *x*, holds a position z in an organization y. Position z has authority. It is the *purpose* of the official *x* to work for the company y and it is the company y that imparts the trust (and the authority) on the official. The official *x* *constitutes* a single individual. An official is *brought into existence* by whatever procedure the company y specifies.
The details of the notation do not matter to the discussion in this section.

case of *Chrysler officials announced* ϕ, which introduces constraints on the content of the complement, restricting it to matters related to the automobile industry in general. *Chrysler officials* is a particularization of *officials* in general and is thus more specific. A more specific source has a more specific domain of expertise (and is thus often more reliable). This means that the lexical coherence between *Chrysler officials* and *announce* is established just like above with the additional constraint for the complement to be coherent with (i.e. to relate to something about) Chrysler.

Lastly, the metonymic use of *Chrysler announced* ϕ causes a violation of the selectional restriction of *announce*, which leads to a coercion of *Chrysler* to an equivalent of *Chrysler officials*[11].

The interpretation process of the three subject NP phrases *Officials, Chrysler officials,* and *Chrysler* in the context of *announce* is entirely due to the lexical semantics of the words involved, and has thus far not involved general commonsense inference, which is one of the main ideas behind GL (namely to minimize commonsense reasoning to those cases that are not systematically related to linguistic phenomena.)

Let us now make the leap to *say*, a reporting verb that has no explicit semantic constraints. We know that anybody can say anything without in fact being of much consequence. The use of *say* as a reporting verb *in the newspaper context*, however, restricts its use to statements where the source has some insight or experience that justifies repeating or rephrasing his/her words. This is due to the relevance conditions associated with the felicitous use of reported speech in an evidential way, as outlined at the beginning of this section. Thus the coherence constraint in the case of *say* is much weaker than in the case of *announce*, because it is inherited from the meta-lexical concept REPORTING–VERB[12] and is not part of the lexical entry itself.

say(A, B)
[Form: REPORTING–VERB]
[Telic: assert(A,B)]

The case of *Officials said* ϕ is analog to the case of *Officials announced* ϕ, with *officials* being in a position of authority or trust filling the requirement associated with

[11] Cf. Section 6.4.2.

[12] See Chapter 7 for details on the semantic class REPORTING–VERB.

the concept REPORTING–VERB. Similarly, *Chrysler officials said that* ϕ derives the basic coherence relation from the concept REPORTING–VERB and constrains the topic of the complement clause to coher with Chrysler on general principles.

Chrysler said ϕ is more marked than *Chrysler announced* ϕ, but indeed does occur frequently, for example in the Wall Street Journal corpus. Interestingly, we find that in the sentences that contain a company name as subject of *say*, the complement phrase is frequently referring back to the subject with the pronoun *it* as in:

(6) a) AEG also said it expects group operating profit to remain at last year's level of 115 million marks.
b) Chantal said it is in advanced stages of testing one of the drugs, Cyoctol, as a topical treatment for mild to severe acne.
c) And Nissan Motor Co., reacting to foreign pressure on Japanese auto makers, said it plans to slash annual vehicle exports in half by the late 1990s.
d) San Miguel said in a report to its stockholders that higher wages, production costs, and interest rates threaten growth prospects for the second half.

This enhances the markedness of the sentences enforcing the same coercion based on a stipulation of metonymy as for *announce*. But note that it is possible to use *say* in its reporting sense with a much weaker coherence relation between the source and the complement than the one required for *announce*.

The coherence relations between three different source NPs, *Officials, Chrysler officials,* and *Chrysler* to two reporting verbs, namely *announce* and *say* here were derived from the lexical semantics of the words involved and we have shown how the compositional behavior of the matrix clause in turn introduces topic constraints on the complement. This means that an issue as inherently pragmatic sounding as restricting the metonymy in subject position of reporting verbs to extensions that bear relevance to the complement clause can *partly* be resolved on the basis of lexical semantics, when treating the lexicon not as a static, passive list of definitions but a dynamic structure that can incorporate procedures that were previously only found in commonsense reasoning systems. The important advantage over pure commonsense reasoning systems lies in the fact that the reasoning in GL is constrained by and limited to syntactic and

lexical semantic knowledge, incorporating conceptual knowledge only where it bears on language behavior.

Now that I have established the basis for deriving the coherence constraints in a Generative Lexicon framework, let me discuss briefly the different uses of these coherence constraints.

The coherence constraint under discussion is essentially a relation with three parameters, the source, the reporting verb, and the complement clause. This relation can be used to resolve ambiguity in one of the parameters, given the other two. Consider:

(7) U.S. News has yet to announce its 1990 ad rates.

U.S. News can refer either to the institution or a particular issue of the newspaper. Both senses are acceptable metonymies with *announce*. The referent can only be resolved based on the coherence constraint between source and complement.

To detect the coherence requires world knowledge about the financing of a newspaper. Disambiguation is therefore not possible based on purely linguistic means. But the coherence constraint can guide the commonsense reasoning very efficiently.

For an example of a metonymy in the complement consider (8):

(8) Warren Winiarski, proprietor of Stag's Leap Wine Cellars in Napa Valley, announced a $75 price tag for his 1985 Cask 23 Cabernet this fall.

To announce a price tag is a violation of selectional restrictions and it is nonsensical. A price tag is itself a medium to announce something, i.e. a price. Thus you can *present* a price tag, *make out* a price tag, but not *announce* a price tag. Whether this is truly a case of metonymy or just bad writing is not of importance in this context. Important is rather that we have no problem understanding the meaning of this sentence as *... announced that his 1985 Cask 23 Cabernet was priced at $75 this fall.* or as *... announced that his 1985 Cask 23 Cabernet would carry a $75 price tag this fall.* This is recoverable from the parameters reporting verb and source in combination with the world knowledge that wine stores fix the price for their merchandise every year. Thus the source representing a wine store and the reporting verb being *announce*, which

indicates that this introduces the *new* price, form a coherence constraint that leads to the correct interpretation.

A case where the correct interpretation of the verb can be recovered through coherence between source and complement is harder to construct, but consider (9):

(9) Then the CEO said blahblahblah.

In this (admittedly contrived) example the missing coherence between source and complement indicates clearly that this is not an evidential use of reported speech, but rather a demonstrative one. Even with this interpretation the example seems awkward and bound to very particular contexts (definitely spoken discourse).

These examples have demonstrated that the coherence constraint guides commonsense reasoning to an efficient disambiguation of any one of the three parameters involved. The task is usually made easier when sentences are considered in the context in which they occur, which will additionally constrain the disambiguation process. It is important, however, to acknowledge that even in isolation a disambiguation can be achieved.

Chapter 6

Corpus Analysis

The last chapter introduced the methodology used in this thesis to derive a lexical semantics for reporting verbs, using corpus analysis and the Generative Lexicon framework. After having introduced some of the advantages of the Generative Lexicon's richer expressive power informally, I will now turn to corpus analysis.

6.1 Usage

There exist two very different semantic devices, for which a computational semantics has to account in different ways.

First, there is *usage*, or conventionalized meaning, as illustrated by idioms and frozen metaphors, as in *"to kick the bucket"* or *"I am freezing"*. Usage is characterized by its change: usage changes over time, as new terms and patterns are coined (compare the omnipresent computer metaphor when talking about mental processes today with usage 100 years ago) and "old" words take on slightly different meanings (cf. [Traugott, 1989, Anderson, 1986] for accounts of meaning shifts towards evidential readings in perceptual and cognitive verbs). But more importantly, usage changes over certain well described contexts; this is part of the *style* of a certain sublanguage, as for example the style of the Wall Street Journal differs from the style of TIME magazine or the style of a mystery novel differs from the style of a romance novel. Of course newspaper style includes both,

the Wall Street Journal and TIME magazine, and is significantly different from the style for "fiction" in general, encompassing both, mystery and romance. It is my firm belief that to recognize the *style*, or *register*, or *sublanguage* is as important for determining the correct meaning for a text, as is the determination of syntactic structure (cf. also [Grishman *et al.*, 1986, Nirenburg and Raskin, 1987]).

Secondly, there is the *compositional meaning*, the composition of several words according to their syntactic requirements, which indicate how their individual definitions combine to the overall meaning. This is the realm of compositional semantics (see also the last chapter)

Both, usage and compositional semantics have found ample attention in the literature; however it is usually the case that semantic theories focus on only one. Linguistics since Chomsky has focused almost exclusively on the compositional aspects of language, a trend which has recently been counterbalanced by computational linguists' interest for the usage of words in particular contexts, made possible by the wider availability of corpora and tools for corpus analysis. Theories that integrate both aspects are currently being developed (cf. [Pustejovsky, 1991]).

Corpus analysis for lexical semantics has in the past suffered from the inadequacy of the available corpora. While very careful work had been done on assembling text fragments from many different sources in order to have a *balanced* corpus that reflects the behavior of the language accross all styles (cf. [Francis and Kučera, 1982]), a one million word corpus such as the Brown corpus does not contain enough occurrences of a single content word to establish its lexical semantics. Thus traditional corpora have been used more successfully to determine syntactic behavior, especially selectional restrictions for verbs based on part of speech tags on the corpus.

Size, however, was not the only shortcoming of traditional, balanced corpora. The very fact that they tried to achieve balance by combining short passages of text from many different sources necessarily had as a result that *contextual* features were not recoverable automatically. But word meaning can only be derived from the contexts, in which the word occurs. Thus again, traditional corpora yielded only very general results.

During the last five years a different kind of corpus has been accessible, a corpus

that does not have to be collected, but can be converted automatically from electronic data carriers used by publishers, newswire services, journals and magazines, etc. These are *homogeneous* corpora, from one source only and therefore have an identifiable *style* and *sublanguage*. The data is abundant (seven million words from the Wall Street Journal from 1989 alone). The sources of these corpora are becoming more varied, so that generalizations over different styles and sublanguages will soon be feasible (the analogue to a balanced corpus). On-line data "relay" stations are being instituted, the Consortium for Lexical Research at New Mexico State University is one example. This sudden availability of data has spurned the creation of many tools to access them and to extract the data automatically.

With the renewed interest in the analysis of large corpora comes an effort to make tools and theories "reusable", that is to standardize the formats for tagging and storing corpora, to describe techniques of data extraction on the implementation level (enabling others to recreate, that is reuse, the techniques), and to devise modular linguistic theories that can be integrated with other modular theories. This thesis is written in the spirit of this new development, see in particular the application neutral representation scheme outlined in Chapter 8.

The following sections introduce some of the current techniques for extracting syntactic and surface collocations from corpora. Section 6.4 introduces the contrasting notion of *semantic collocation* and contrasts some results of a semi-automatic corpus analysis with the limitations of a fully automatic approach to deriving lexical semantics.

6.2 Current Methods

Many current efforts in corpus analysis in computational linguistics aim to automatize the techniques applied to traditional corpus analysis, e.g. part of speech tagging [Church, 1988]. I will not discuss the literature on these basic corpus analysis tools.

The work presented in this section is based on the notion of *cooccurrence*, the fact that two words occur within a certain distance from each other (the default is cooccurrence within a sentence). Probabilistic measures of how likely the two words cooccur are derived from the corpus using statistical notions such as *mutual information*, and

probability.

6.2.1 Mutual Information

Mutual information is a notion from information theory [Fano, 1961]. There, mutual information is defined as a measure of the interdependence of two events in sequence.

Given $x \in X$, $y \in Y$ are independent events, then

$$\mathcal{P}_{X,Y}(x,y) = \mathcal{P}_X(x)\mathcal{P}_Y(y)$$

Given $x \in X, y \in Y$ are dependent events, then

$$\mathcal{P}_{X,Y}(x,y) = \mathcal{P}_X(x) \geq \mathcal{P}_X(x)\mathcal{P}_Y(y)$$

Given $x \in X, y \in Y$ are mutually exclusive events, then

$$\mathcal{P}_{X,Y}(x,y) = 0$$

Mutual Information is then defined as:

$$\mathcal{MI}(x.y) = log_2 \frac{\mathcal{P}_{X,Y}(x,y)}{\mathcal{P}_X(x)\mathcal{P}_Y(y)}$$

The notion of mutual information was used by Magerman and Marcus [Magerman and Marcus, 1990] to build a *distituent parser*. The distituent parser operates on the idea that certain syntactic categories cannot be adjacent in a sentence constituent. If two categories that cannot be adjacent in a constituent occur next to each other in a sentence, they trigger a constituent boundary. Magerman and Marcus derive what syntactic categories cannot be adjacent through a mutual information analysis on the tagged Brown corpus. The notion of mutual information was extended to consider not only mutual information between two words, but mutual information for *n-grams*, n words in sequence. The algorithm is augmented with a small distituent grammar, containing well known constituent boundary cases such as *noun prep* for English.

6.2.2 Association Ratios

Church and Hanks [Church and Hanks, 1990] modify the notion of mutual information to allow for different restrictions on the environment in which cooccurrence is defined. Cooccurrence in their system is only significant if the words appear within a certain distance from each other. They introduce the notion of a window around a key word; cooccurrence is then only defined within this window. The size of the window is variable and the keyword does not have to lie in the middle of the window[1]. This measure is called *association ratio* and is determined by:

$$\mathcal{I}_w(word_1, word_2) = log_2 \frac{cooccurrence_w(word_1, word_2) \times N}{occurrence(word_1) occurrence(word_2)},$$

where N is the size of the text and w indicates the window size and direction.

Church and Hanks apply this measure to investigate lexico-syntactic relationships between verbs and typical arguments and adjuncts. They use the measure to determine phrasal verbs (based on [Sinclair, 1987]) such as *set up, set off, set out*, refer to Hindle's work reported below, and suggest association ratios as a tool for lexicographers to find frequent word cooccurrences such as *save from*, which they cite as an example of a cooccurrence that is significant (21 times higher than chance) but not frequent enough to be necessarily evident. The lexicographer can also use the association ratios to infer *semantic categories*, e.g. from the fact that *save* cooccurs frequently with *month* and *annually* a word sense that relates *save* to time can be inferred, that might otherwise go unnoticed.

6.2.3 Similarity Metric for Predicate Argument Structures

Hindle [Hindle, 1990] applies the mutual information measure to the output of his parser Fidditch [Hindle, 1983]. Fidditch is a robust, non-committal parser that will output partial parses for sentences it cannot parse correctly. Hindle computes the mutual information measure for the predicate and one of its arguments, for example the weighted list of objects for *drink* in his corpus was *bunch beer, tea, Pepsi, champagne, ...* Based on the distributional hypothesis that the degree of shared contexts is a

[1] This of course means that *association ratio* is not a symmetric relation.

similarity measure for words, he develops a similarity metric for nouns, SIM_{subj} and SIM_{obj} as the minimum shared cooccurrence weights (mutual information).

Object and subject similarity:

$$SIM_X(v_i n_j n_k) = \begin{cases} min(C_X(v_i n_j), C_X(v_i n_k)), & if C_X(v_i n_j) > 0 \text{ and } C_X(v_i n_k) > 0 \\ abs(max(C_X(v_i n_j), C_X(v_i n_k))), & if C_X(v_i n_j) < 0 \text{ and } C_X(v_i n_k) < 0 \\ 0, & otherwise \end{cases}$$

where X = *subj or obj.*

Noun similarity:

$$SIM(n_1 n_2) = \sum_{i=0}^{N} SIM_{subj}(v_i n_1 n_2) + SIM_{obj}(v_i n_1 n_2)$$

Hindle thus finds sets of semantically similar nouns based on syntactic cooccurrence data. The sets he extracts are promising; for example, the ten most similar nouns to *treaty* are in his corpus: *agreement, plan, constitution, contract, proposal, accord, amendment, rule, law, legislation.*

6.2.4 Collocational Constraints

Smadja and McKeown [Smadja and McKeown, 1990] identify collocational patterns of two or more words and represent them for use in a generation system. They consider three main classes of collocational patterns, *open compounds*, that is uninterrupted sequences of words, *predicative relations*, that is two or more words often used together in a similar syntactic relation, and *phrasal templates*, that is idiomatic phrases that can possibly contain one or more empty slots. Xtract, developed by Smadja, extracts the different patterns in two stages, first building a concordance and deriving statistics from that, then in stage two iterating the process, starting with the word pairs found in stage one. The statistical techniques used are different from the previously mentioned ones, but are not detailed in [Smadja and McKeown, 1990].

6.3 A Filtering Algorithm

The statistical data used in this dissertation have been extracted using the techniques described in the previous section. The focus was on exploring the *semantic collocations* for reporting verbs (see following section), which can not be extracted fully automatically. Aside from individual routines that extract full sentences containing a keyword from the corpus or particular windows around such keywords, the most useful tool proved to be a filtering algorithm that combined all these techniques.

I will present the idea behind this algorithm here without giving the actual code and show some results derived with the algorithm.

The algorithm proceeds roughly in the following steps:

Extract environment: Search the corpus for all occurrences of the key and extract its immediate environment, x words to the left and y words to the right.

Sort and count: Sort and count all words extracted in the previous step. Extract the n most frequent words among them.

Mark possible nouns: For the n most frequently cooccurring words, do a dictionary lookup and extract the words that could possibly be nouns.

Determine association ratios: For the set of possible nouns, determine the word's total occurrence in the corpus and the association ratios between the nouns and the key.

Let me first introduce the subprocedures used in the algorithm. EXTRACT is the basic procedure that extracts the immediate context around a keyword. The actual procedure has many different modes and is written in C for efficiency. For illustrative purposes I include a simplified version.

EXTRACT(key, l, r) ; result is a list of words

$\forall$ sentences s in the corpus
  $\forall$ words $w_i \in s$
    if $w_i = key$

then $\forall\ w_j \in s$ such that $i - l \leq j \leq i + r$
accumulate w_j

Note that w_j is a word with position j in the sentence. The sentence boundaries are absolute boundaries, i.e. windows do not cross sentence boundaries.

ASSOCIATION-RATIOS computes the association ratios as described in Section 6.2.2.

ASSOCIATION-RATIOS(Co, $Oc1$, $Oc2$, N)

return $\log_2 \frac{Co \times N}{Oc1 Oc2}$

A third subprocedure filters a list of words according to their possible part of speech. DICTIONARY-LOOKUP($list$, NOUN) returns the list of words that have a noun entry in the dictionary. Note that this does not guarantee that the word had been used with this part of speech in the corpus. The dictionary used is an on-line version of the Oxford Advanced Learner's Dictionary [Hornby, 1974].

DICTIONARY-LOOKUP($list$, $part\_of\_speech$) ; result is a list of words

```
∀ words w_i ∈ list
    if   has-word-sense(w_i, part_of_speech)
    then accumulate w
    else if   last-letter(w_i) = "s"
         then w_i' ← w_i without final "s"
              if   has-word-sense(w_i', part_of_speech)
              then accumulate w
         else if   last-letters(w_i) = "ed"
              then w_i' ← w_i without final "ed"
                   if   has-word-sense(w_i', part_of_speech)
                   then accumulate w
```

The procedure DICTIONARY-LOOKUP takes as arguments a list of words and a part of speech, returning the subset of the words of the first argument that has as a possible word sense the part of speech specified. If lookup fails and the word ends in an "s" then the procedure looks up the "root" of the word without the final "s". If the word ends in "ed", the procedure tries the word without the last two letters. Note that the

dictionary table contains irregular verb forms. The procedures are not case sensitive.

The overall algorithm can then be described as follows:

FILTER(*key*, *l*, *r*, *pos*)

> $N \leftarrow$ number of words in *corpus*
> *window_words* $\leftarrow$ extract(*key*, *l*, *r*)
> *unique_words* $\leftarrow$ all unique words in *window_words*
> *selected_words* $\leftarrow$ dictionary-lookup(*unique_words*, *pos*)
> *top_words* $\leftarrow$ 500 words in *selected_words* most frequent in *window_words*
> $\forall$ words $w \in$ *top_words*
>
> > *total* $\leftarrow$ count(w, *corpus*) ; total occurrence of w in *corpus*
> > *cooc* $\leftarrow$ count(w, *window_words*) ; cooccurrence of *key* with w
> > *key_total* $\leftarrow$ count(*key*, *window_words*) ; total occurrence of key in *corpus*
> > output w and association-ratio(*cooc*, *key_total*, *total*, N)

This algorithm has been implemented and runs fully automatically. It is a useful tool to explore the possible space of lexicalized characteristics of a word in an untagged corpus. However, as described this algorithms may yield a set up to 500 words[2] plus their statistics, most of which are clearly not used as either subject or object (a recurring "red herring" is *in*: the OED lists one word meaning for *in* as a noun and thus *in* occurs in virtually every result from the above algorithm).

One obvious way to further process the output is as suggested in the literature: prune all the words that cooccur less than 6 times with the key [Smadja and McKeown, 1990]. This cuts down the number of words significantly.

Figure 6.1 shows the result for a search for possible nouns cooccurring within the same sentence as *dispute*, ordered by association ratios. The interesting items in this list are *resolve, bitter,* and *settle*, which are clearly not used as nouns in the context of *dispute*, but show phrasal patterns: *bitter disputes* are either *resolved* or *settled*. Likely nouns of interest are *settlement, patent, labor,* and *court*, none of which are likely to occur in subject position, but can occur as compound nominals or complements: *labor dispute*, *dispute about settlements*, etc.

[2]The number is actually parameterized and can be changed. Tests run with a setting of 500 took roughly three hours on a medium power workstation.

Key	co-oc	total	assoc. ratio	Key	co-oc	total	assoc. ratio
dispute		584		oil	8	2867	5.0640
resolve	22	348	9.5656	trade	15	6963	4.6906
bitter	10	193	9.2787	government	10	6342	4.2404
settlement	17	791	8.0092	tax	8	5896	4.0238
patent	7	354	7.8889	most	10	7603	3.9788
resolution	6	314	7.8396	company	22	17241	3.9352
centers	6	331	7.7635	long	6	5821	3.6270
settle	19	1498	7.2482	last	8	8419	3.5096
ruling	7	649	7.0144	two	9	9641	3.4843
labor	19	1994	6.8358	even	7	11212	2.9039
stems	9	1925	5.8087	million	11	25677	2.3604
over	94	23550	5.5805	one	11	28420	2.2141
contract	13	3434	5.5041	are	21	56141	2.1649
court	14	3726	5.4931	in	150	529781	1.7631
claim	6	1978	5.1844	year	6	26334	1.4497
issues	8	2705	5.1479	no	10	62768	0.9332

Figure 6.1: Possible nouns cooccurring with *dispute* in WSJC

The output of the algorithms does not provide any "results". The apparent failure rate of this algorithm is partly due to the dictionary used to determine possible part of speech for words: The limited, general vocabulary of the Oxford Advanced Learner's Dictionary on-line version will miss a large set of true nouns, including names and titles (*Fitzwater, CEO*). However, some frequent source descriptions, for instance, can be recognised when comparing the output of the algorithm for different verb keys; likely sources in Figure 6.1 are *government* and *company*.

This algorithm has been provided to illustrate the methodology applied in detail and to show just how much "semantic" information can be extracted from a corpus automatically without using a parser or tagged text (the hit rate of the algorithm is dramatically increased when using a tagged corpus rather than dictionary lookup.) The current statistic techniques (association ratios, cooccurrence counts, etc.) serve to *prepare* the data for the human analyst, not extract them. For research purposes I prefer this semi-automatic method — it has given me much insight in numerous runs through the algorithm.

6.4 Semantic Collocations

The methods described in the previous sections all rely on the collocation of specific words or syntactic categories. The results have been used for parsing, generation, and lexicographic research. Only Hindle's effort can be seen as a step towards defining the lexical semantics of words. Hindle, too, however falls short of considering truly semantic properties. He fails, for instance, to make the generalization that the nouns found as objects of *drink* are all *liquids*, in fact a selectional restriction that *drink* imposes.

The focus of current research on surface phenomena stems from the goal to fully automate lexical semantic discovery procedures. Automatic approaches are, however, limited to extract exactly the information they have been designed to extract. Generalizations over the results are often not possible, when the result is encoded in a measure such as association ratios. Here I argue to design the statistical analysis tools in a way that allows the human analyst to discover data that is not readily available to an automated analysis. In particular, I will explore the notion of *semantic collocations* in this section.

Semantic collocations are semantic generalizations over cooccurrence data, restricted to those connotations that are not selectional restrictions in the traditional sense.

Semantic collocations span a wide variety of phenomena. It is the set of subtle, conventionalized semantic collocations that distinguishes the language of a non-native speaker from a native speaker's, that distinguishes sublanguage uses, and that even distinguishes between the classification of a violation of selectional restrictions as a logical metonymy, a novel metaphor, or simply a mistake (see Section 6.4.2 below for more detail).

The next sections will identify several very different semantic collocations and point out their importance for lexical semantics in general and the semantic field of reporting verbs in particular.

6.4.1 Discourse Polarity Items

It is possible to acquire information concerning lexical presuppositions and preferences from corpora, when analyzed with the appropriate semantic tools. In particular, I will discuss here the phenomenon of *discourse polarity*, and how corpus-based experiments provide clues towards the representation of this phenomenon, as well as information on preference relations.

Negative polarity items are words such as *any*, which have to occur within the scope of a negation[3]. It appears that verbs, too, can specify for negative polarity, but in a looser form, here called *discourse polarity*. *Deny* is an instance of a reporting verb that specifies in its definition for an opposition. It is interesting to observe that this negative polarity in the complement of *deny* is strong enough to allow for negative polarity items:

(1) He denied any wrongdoing.

An even more interesting case than *deny* is *insist*. *Insist* does not denote an opposition, as does *deny*. But *insist* still presupposes an opposition to be recoverable from the immediate context. To illustrate this, consider the following sentences taken from the Wall Street Journal corpus:

(2) (a) But Mr. Fourtou insisted that the restructuring plans hadn't played a role in his decision.
(b) But so far, the majority is insisting that a daily paper in the home is an essential educational resource that Mr. Oshry must have, like it or not.
(c) But Mr. Nishi insists there is a common theme to his scattered projects: to improve and spread personal computers.
(d) "Mister, Djemaa is a crazy place for you," insists the first of many young men, clutching a visitor's sleeve.
(e) But the BNL sources yesterday insisted that the head office was aware of only a small portion of the credits to Iraq made by Atlanta.
(f) Mr. Smale, who ordinarily insists on a test market before a national roll-out, told the team to go ahead – although he said he was skeptical that Pringle's could survive, Mr. Tucker says.
(g) The Cantonese insist that their fish be "fresh," though one whiff of Hong Kong harbor and the visitor may yearn for something shipped from distant seas.
(h) Money isn't the issue, Mr. Bush insists.

[3]There is a rich literature on this topic. For discussion, see [Ladusaw, 1980, Leinbarger, 1980]

From analyzing these and similar data, a pattern emerges concerning the use of verbs like *insist* in discourse; namely, the cooccurrence with discourse markers denoting negative affect, such as *although* and *but*, as well as literal negatives, e.g. *no* and *not*. The opposition is expressed in form of plain negation in the complement clause, as in Examples (2a) and (2h), through discourse markers such as *but, only,* or *ordinarily* (Examples (2a), (b), (c), (e) and (f)), through adjunct clauses, as in Examples (2b) and (g). Only in very few complements to *insist* do we not find a marker of *discourse polarity*, but have to construe the opposition from the surrounding context, as in Example (2d).

The lexical definition of *insist* in the Longman Dictionary of Contemporary English (LDOCE) [Procter, 1978] is

insist **1** to declare firmly (when opposed)

and in the Merriam Webster Pocket Dictionary (MWDP) [Woolf, 1974]:

insist to take a resolute stand: PERSIST.

The opposition, mentioned explicitly in LDOCE but only hinted at in MWDP, is not part of the decompositional structure of *insist*. This is indicated by the parentheses in LDOCE and the fact that the opposition is only indirectly implied in MWDP. The opposition, we have to assume, is *presupposed*.

That hypothesis was confirmed on the seven million word corpus of Wall Street Journal articles from 1989 (WSJC). An analysis of the pattern of cooccurrence of certain markers of opposition with *insist* yielded the results shown in Figure 6.2:

The cooccurrence rate of *insist* with markers of discourse polarity is strong enough to characterize *insist* as a discourse polarity item.

We can summarize the discussion of the lexical semantics of *insist* with a preliminary definition, making explicit the underlying opposition to the assumed context (here denoted by ψ) and the fact that *insist* is a reporting verb.

(3) **insist(A,B)**
[Const: MANNER: vehement]
[Form: REPORTING–VERB]
[Telic: assert(A,B) & opposed(B,ψ)]

[Agentive: human(A)]

Keywords	Occ	Comments
insist	586	occurrences throughout the corpus
insist on	109	these have been cleaned by hand and are actually occurrences of the idiom *insist on* rather than accidental cooccurrences.
insist & but	117	occurrences of both *insist* and *but* in the same sentence
insist & negation	186	includes *not* and *n't*
insist & subjunctive	159	includes *would, could, should,* and *be*
insist & but & neg.	14	
insist & but & on	12	
insist & but & subj.	8	

Figure 6.2: Negative markers with *insist* in WSJC

This states that the telic role for a verb such as *insist*, is an "assertion", where there is a contextually introduced proposition, here represented by ψ, that stands in opposition to that which is asserted (i.e. B). As argued in [Pustejovsky, 1991, Miller and Fellbaum, 1991], such simple oppositional predicates form a central part in the lexicalization of concepts. Semantically motivated collocations such as these extracted from large corpora can provide presuppositional information for words which would otherwise be missing from the lexical semantics of an entry. The importance of capturing these presuppositions is that they have to be "fulfilled" within the text and therefore indirectly guide the interpretation process: if the presupposed opposition is not expressed directly (within the sentence or surrounding co-text), a common sense reasoning process has to derive it (examples for such "common sensical" oppositions are *Republican—raising taxes* and *Democrat—cutting social services*, or simply *Republican—Democrat* in the U.S.).

The importance of marking discourse polarity items lies in the fact that they provide

another kind of coherence constraint that is triggered lexically. To obey this coherence constraint is important for generation systems in order to produce good text. And as shown in the last chapter, coherence constraints help in text understanding in general. The particular importance of discourse polarity items among reporting verbs, however, is that they influence the evaluation of the *reliability* of the reported statement.

Chapter 3, Section 3.1.1, described Givón's partitioning of the epistemic space into three major categories. The category of highest certainty contains propositions that are unchallangeable and do not require evidential justification. The category of medium certainty contains propositions that can be challenged by the hearer/reader and require evidential support. The category of lowest certainty contains propositions that are of hypothetical nature or otherwise beneath challenge and evidential support. While reported speech in newspaper articles falls into the category of medium certainty according to my analysis, the reporter can indicate his evaluation by choosing the appropriate reporting verb. In choosing *announce*, a factive verb requiring the source to have some legitimation to speak on the topic, the reporter indicates that he or she attributes high certainty to the statement. By choosing *insist*, on the other hand, the reporter points out that the statement is not only open to challenge but has in fact been challenged. This necessarily lowers the certainty of the reported statement for the reader (unless the reader has some superior knowledge). All discourse polarity items then have that same effect of pointing out an opposition to the reported statement, thus requiring additional substantiation in order to be considered reliable.[4]

Let us assume an interpretation system that classifies knowledge gained from texts as *above, open* (to), and *below* challenge. According to Givón's scheme this system will place knowledge that was presupposed in the text or shared knowledge of the universe as coded in the lexicon into category *above*, realis assertions in the category *open*, and irrealis assertions in the category *below*. This first step of categorization is syntactic-semantic based. The instructions to the commonsense reasoning system resulting from this classification are sketched in the following pseudocode procedure:

$\forall$ i $\in$ information
    if category(i) = above

[4]Section 7.2.2 in the following chapter will introduce more such *semantic dimensions* that influence directly or indirectly the evaluation of the reliability of reported speech.

```
then if     new-information(i)
     then enter-as-background-knowledge(i)
     else if     ¬consistent-with-knowledge-base(i)
          then do-research-on-this-issue(i)
else if     category(i) = open
     then evaluate-evidential-support(i)
     else if     category(i) = below
          then annotate-with-context-variables(i)
               store-as-hypothetical(i)
```

This procedure instructs commonsense reasoning to accept (and incorporate) information of the category *above* if it is new information to the knowledge base. If the knowledge base already contains conflicting information, the issue is to be researched further. The commonsense procedure will then study back issues of the newspaper and try to establish evidence for either of the two conflicting pieces of knowledge. The details of this research step are beyond the scope of this discussion.

Information of the category *open*[5] is subject to the subprocedure EVALUATE-EVIDENTIAL-SUPPORT. Reported speech is always part of the category *open* and the subprocedure contains of course the evaluation step discussed in Chapter 3. The general evaluation procedure can not be outlined here, let me instead focus on reported speech and in particular the role of discourse polarity items in this procedure. Note that the notation is the same as in Chapter 3, where Ri:OC:U refers to the reporting verb and Ci refers to the primary information. DPI refers to the set of discourse polarity items:

EVALUATE-EVIDENTIAL-SUPPORT(i)

```
if     reported-speech(i)
then if     Ri:OC:U ∈ DPI
     then find-scope-of-opposition(Ci)
          if     category(opposition(Ci)) = above
          then reliability(Ci) := low
          else if     category(opposition(Ci)) = below
               then reliability(opposition(Ci)) := low
                    reliability(Ci) := moderate
               else if     evaluate-on-personal-knowledge(Ci) >
```

[5]Information in this category is usually in form of propositions.

evaluate-on-personal-knowledge(opposition(Ci))
then reliability(Ci) := moderate

This example is a very crude version of a possible evaluation function that uses the (lexical semantic) knowledge of discourse polarity items. In summary this procedure tries to evaluate an instance of reported speech Si = Ci(Ri). Upon detecting that the reporting verb is a discourse polarity item the procedure calls a subprocedure that determines the scope of the opposition (and its content). The evaluation of Si depends on the evaluation of the opposition to Ci. Is the category of the opposition *above* (challenge), the reliability of Si is deemed *low*[6]. Is the category of the opposition to Ci *below* (challenge), the reliability of the opposition is deemed low and the reliability of Ci is deemed *moderate*. Is the category of the opposition to Ci *open* (to challenge), that is both Ci and its opposition have the same status in this respect, then the reader evaluates both Ci and its opposition based on personal background knowledge. If Ci fits better in the personal knowledge base of the reader than its opposite, then Ci is assigned a *moderate* reliability (*moderate* is chosen here because no true evidence was supplied in the text as considered here. Note that this is not a realistic situation; other coherence relations and the surrounding context will usually supply more information.)

While this example of a possible evaluation procedure for discourse polarity items in isolation does not give any insights into its impact on text understanding in general, it still demonstrates that discourse polarity is a powerful notion even in the absence of context.

6.4.2 Metonymy

The previous section focused on semantic collocations as part of the lexical semantics of a word. This section demonstrates a more elusive type of semantic collocations, namely the preference of a verb to allow metonymy in different argument positions.

Metonymy is a field of active research (cf. [Fass, 1988, Anick and Bergler, 1991]). Different researchers define metonymy slightly differently; the working definition I am assuming here characterizes metonymy as a licensed violation of selectional restrictions

[6]The three values *high, moderate,* and *low* stand here in lieu of a true evaluation.

in argument position. Metonymy is similar to metaphor, and indeed some researchers do not distinguish between metaphor and metonymy. Metonymy as defined here is not as freely applicable as metaphor — the only constraint on using something metaphorically is that the hearer or reader can understand the metaphor. Metonymy is, in contrast, *conventionalized*, yet not ridgid or "frozen". Examples are:

(4) **Extended metonymy:**
   a) My car drinks gasoline. [Wilks, 1978]
   b) I'm the ham sandwich; the quiche is my friend. [Fauconnier, 1985]

(5) **Logical metonymy:**
   a) Mary enjoyed the book. [Pustejovsky, 1991]

   b) The government has repeatedly refused to deny that Prime Minister Margaret Thatcher vetoed the Channel Tunnel at her summit meeting with President Mitterand on 18 May, as *New Scientist* revealed last week. [New, 1982, Hobbs, 1985]

The fact that metonymy as I am defining it here is *conventionalized* means that it is recoverable only from *usage*, and is thus dependent on corpus analysis. The following sections show how many facets there are to this single phenomenon which have to be understood before an attempt to automate the process can be made.

Logical Metonymy

Although metonymy is a general device — in that it can appear in almost any context and make use of associations never considered before — a closer look at the data reveals, however, that metonymy as used in newspaper articles is much more restricted and systematic, corresponding very closely to *logical metonymy*.

Pustejovsky [Pustejovsky, 1991] defines a subset of metonymic phenomena as *logical metonymy*. For him, metonymy exists where a subpart or related notion of an object "stands for" the object itself. *Logical metonymy* is defined to occur when a logical argument (i.e. subpart) of a semantic type that is selected by some (contextually determined) function, denotes the semantic type itself. Thus logical metonymy is based on a relation recoverable from the lexical structures of a generative lexicon, where general

metonymy may rely on world knowledge to link the description to the target referent. Logical metonymy can be resolved by *type coercion*, "a semantic operation that converts an argument to the type which is expected by a function, where it would otherwise result in a type error" [Pustejovsky, 1991]. An example for logical metonymy and type coercion is *Mary enjoyed the book*, where *enjoy* selects for a process, which coerces *the book* to be reinterpreted as a default process (or event) associated with the lexical item *book*, here found in the telic role of *book*, namely *read*.[7]

An example of a typical metonymic extension found with reporting verbs is *synechdoche*, where the whole stands for a part. For example, while the verb *announce* selects for a human subject, sentences like *The Phantasie Corporation announced third quarter losses* are not only an acceptable paraphrase of the selectionally correct form *Mr. Phantasie Jr. announced third quarter losses for Phantasie Corp*, but they are the *preferred* form in the corpus being examined (i.e. the Wall Street Journal). This is an example of subject *type coercion*, as discussed in [Pustejovsky, 1992]. For example, the qualia structure for a noun such as *corporation* is given below:

corporation(*x*)

[Const: GROUP(people(y)), spokesperson(w), executive(z)]
[Form: organization(*x*)]
[Telic: execute(z,decisions)]
[Agentive: incorporate(y,*x*)]

The metonymic extension in this example is straightforward: a spokesman, executive, or otherwise legitimate representative, "speaking for" a company or institution can be metonymically replaced by that company or institution.[8]

We find that this type of metonymic extension for the subject is natural and indeed very frequent with *reporting verbs* [Bergler, 1991], such as *announce, report, release, claim, etc.* while it is in general not possible with other verbs selecting human subjects; e.g. motion verbs (with the exception of *move* and *go* and metaphoric use); or verbs

[7]This sketch of the theory of coercion is necessarily oversimplified. See [Pustejovsky, 1991, Pustejovsky, 1995, Pustejovsky, 1992] for the presentation of the issues and solutions involved.

[8]Note, however that the metonymic extension is not quite as simple as extending from any *employee* to the whole company or institution, but that a form of *legitimation* has to be involved. This constraint can be met by a form of co-specification.

of contemplation (such as *contemplate, consider, think*). However, there are subtle differences in the occurrence of such metonymies for the different members of the same semantic verb class, which become evident from corpus analysis.

6.4.3 Lexicalized Preference for Metonymy

Reporting verbs with a clausal complement display metonymy only in subject position. The data show that metonymy is frequent in subject position with reporting verbs. A closer analysis reveals, however, that different types of logical metonymic extensions are accepted with differing frequency by individual reporting verbs. More precisely, they seem to distinguish between a single person, a group of persons, and an institution.

A careful study of seven reporting verbs on a small, 250,000 word corpus of TIME magazine articles from 1963, found that the preference for different metonymic extensions varies considerably within this field. Figure 6.3 shows the findings for the words *insist, deny, admit, claim, announce, said* and *told* for two metonymic extensions, namely where a *group* stands for an individual (*Analysts said ...*) and where a *company* or other *institution* stands for the individual (IBM announced ...) (cf. [Bergler, 1991]).

	person	group	instit.	other
admit	64%	19%	14%	2%
deny	59%	11%	19%	11%
insist	57%	24%	16%	3%
said	83%	6%	4%	8%
told	69%	7%	8%	16%
announce	51%	10%	31%	8%
claim	35%	21%	38%	6%

Figure 6.3: Preference for different metonymies in subject position

The differences in patterns of metonymic behavior are quite striking: semantically similar verbs seem to pattern similarly over all three categories; *admit, insist*, and *deny* show a closer resemblance to each other than to any of the others, while *said* and *told* form a category by themselves. A purely semantic explanation why *said* and *told* seem to not prefer the metonymic use in subject position could be construed, conjecturing that these verbs relate more closely to the act of uttering, or that they are stylistically

too informal. Evidence from other corpora, however, suggests that such information is accurately characterized as lexical preference. A comparative analysis of a subset of the Wall Street Journal Corpus, for example, shows that *said* has a quite different metonymic distribution there, reported in Figure 6.4.

corpus size: 160,000
occurrences of *said*: 914

	person	group	institution	other
total in WSJ	449	135	311	19
WSJ in %	49%	15%	34%	2%
TIME in %	83%	6%	4%	8%

Figure 6.4: Preference for metonymies for *said* in a fragment of the Wall Street Journal corpus.

In the Wall Street Journal corpus *say* selected for an individual person in only 50% of the sentences, while a company/institution appeared in subject position in 34% of the cases.

It is interesting to note here that the distribution of metonymic preference for *say* in WSJC and for *announce* in the TIME corpus is very similar and that the usage of *say* in the Wall Street Journal closely resembles the definition of *announce* in that only legitimized sources are quoted on topics that fall well into their expertise. This observation supports the hypothesis that the difference in distribution for *say* in WSJC and the TIME corpus is not accidental but constitutes a style difference, reflecting different semantics in the two corpora. The frequent use of metonymic extension form persons to companies and institutions in the Wall Street Journal is due to its particular focus on the business world and politics.

'Government' Metonymy

The TIMEcorpus also exhibited a preference for one particular metonymy, which is of special interest for reporting verbs, namely where the name of a country, of a country's citizens, of a capital, or even of the building in which the government resides stands for the government itself. Examples are *Great Britain/ The British/ London/ 10 Down-*

ing Street announced.... This metonymy is of great importance to parsers, because it deviates crucially from the standard metonymic use. Consider the example *Washington announced tax cut-backs*. We assume *Washington* stands for the U.S. government. This overrides the default metonymic assumption at work in examples like: *Waltham announced increased revenues*, where *Waltham* stands for the city of Waltham.

Verb	percent of all occurrences
admit	5%
announce	18%
claim	25%
denied	33%
insist	9%
said	3%
told	0%

Figure 6.5: *Country, countrymen,* or *capital* standing for the *government* in subject position of 7 reporting verbs.

Figure 6.5 shows the preference of the reporting verbs for this metonymy in subject position. Again the numbers are too small to say anything about each lexical entry, but the difference in preference is strong enough to suggest it is not only due to the specific style of the magazine, but that some metonymies form strong collocations that should be reflected in the lexicon. Such results in addition provide interesting data for preference driven semantic analysis such as Wilks' [Wilks, 1975].

6.5 Summary

This chapter presented current corpus analysis methods and a particular algorithm that combines these methods in an exploration tool that can help a human analyst to detect *semantic collocations*. These semantic collocations can not be extracted automatically, in fact not much research has been done in systematically identifying semantic collocations. To show their importance for text understanding and their impact on constraining commonsense reasoning, two examples of different semantic collocations have been provided, namely the property of certain reporting verbs to cooccur with an opposition

expressed in the immediate context (but not necessarily within the sentence). I have called these reporting verbs *discourse polarity items*. Another important semantic collocation is the differing preference of reporting verbs for different kinds of metonymy in subject position. To represent preferences for metonymy lexically can significantly reduce inference time. The findings of the corpus analysis show that this preference is not homogeneous throughout the field of reporting verbs and therefore not a commonsense meta-rule over the whole field but must be established for every word individually. That the results seem to further partition the field of reporting verbs into finer semantic classes is a beneficial side-effect that will lead to a better understanding of this field.

Chapter 7

Lexical Entries

This chapter combines the techniques and ideas developed in the previous chapters in deriving the lexical semantics of a set of important reporting verbs and source descriptions. A formalization of the functional role of reported speech in the Generative Lexicon framework is given, introducing a new device, called *semantic class*, that associates essentially *pragmatic information* (about the underlying situation and function in newspaper articles) with *Lexical Conceptual Paradigms*, LCPs. Section 7.1 introduces the terminology and derives the semantic class REPORTING–VERB. Section 7.2 describes the semantic field of reporting verbs and presents the lexical entries for a set of reporting verbs. Section 7.4 derives the lexical entries for a set of frequent source descriptions with special emphasis on the implied trustworthiness.

7.1 Semantic Class

Chapters 3 and 4 motivated the evidential analysis of reported speech in newspaper articles based on the fact that this phenomenon has a clear function in that context. The evidential analysis links linguistic information (i.e. reported speech) with information about the underlying situation, which enables the common sense inferencer to assign credibility to the reported statement.

In order to link pragmatic information to lexical entries, *meta-lexical* structures

have to be described. The Generative Lexicon already provides Lexical Conceptual Paradigms, LCPs, which link syntactic structures to semantic values. Here I introduce a second level of meta-lexical information, linking LCPs to pragmatic information, called *semantic classes*. Semantic classes, like LCPs, provide default information that is inherited by the lexical entries. Lexical entries have to specify their semantic class.

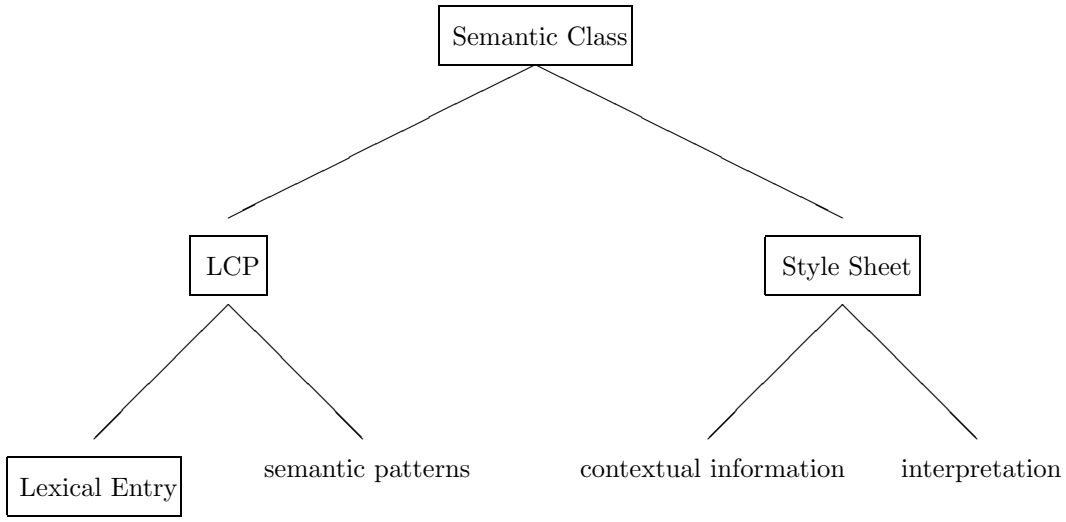


Figure 7.1: Meta-lexical structure.

The reporting verbs considered here form what I will call a *functional field*, that is they are related semantically and can perform the same function, namely to report someone's utterance (in the context of newspaper articles[1]). Thus a functional field is related to a semantic field; in fact it is a specific type of semantic field (cf. [Trier, 1934, Porzig, 1934, Miller and Johnson-Laird, 1976]).

The functional field will be defined by a central, abstract *semantic class* REPORTING–VERB, consisting of an LCP REP–VERB and a *Style sheet* NEWS–REPORTING.

[1]That is in a somewhat formal context. While *"Run, run, the bad wolf is coming" oink-oinked the little pig.* might be acceptable in a narrative for children, *oink-oink* is not part of the functional field under consideration here.

7.1.1 LCP REP–VERB

An LCP links the lexical entry to a mapping from (conventionalized) syntactic patterns to semantic interpretation. Thus it is an LCP that links transitive verbs to their unaccusative forms or which specifies *figure/ground* alternations in certain nominals, etc. (cf. [Pustejovsky, 1991]).

Here I introduce LCPs linking the syntactic components of reported speech sentences to the evidential scope introduced in Section 3.4 on page 52. The notation is as follows:

(1) **LCP**

REP–VERB(V,x,y)

V(T:transition, x:human, y:proposition) $\Longrightarrow$
y(CC,OC,TC),
where [OC:(S=x, U=V)].

Verb V (which specifies for the LCP REP–VERB) is of event type T (a transition for reporting verbs) and has arguments x (external argument — a human subject by default) and y (complement clause of type proposition) and maps into the evidential scope such that y(CC,OC,TC) has been asserted with OC being partially defined as S=x (source is subject) and U=V (utterance characterized by reporting verb). Note that this is minimal default information — if there is additional modifying material in the reporting clause that further specifies the source S or the utterance U, this must be added to the text representation.

LCPs can be specified in the lexical entries directly or they can be subsumed by a semantic class, as is the case for reporting verbs. Because the LCP connects syntactic structures to semantic patterns, it influences the compositional semantics. REP–VERB is an example of an LCP that changes control in the interpretation process, making the complement clause into a first class object, according to the evidential scope.

7.1.2 Style Sheet NEWS–REPORTING

The particular style in which newspaper articles make use of different lexical entries cannot be reflected in the general entry of that word or group of words. The LCP, which is capable of linking syntactic structures to pragmatic information, is also not the right place for style information, because newspaper style changes and can differ significantly. I introduce therefore a new device, *style sheet*, which holds the pragmatic default knowledge for newspaper reporting[2], that summarizes the function of reported speech in newspaper articles:

(2) **Style sheet**

NEWS–REPORTING(V,S,O')

[Context: reporter(R), source(S), utterance(O), situation(UC):
utter(S, O), witness(R, O), interpret-as(R, O, V(S,O'))]
[Commitment: is-committed(R, consistent(UC, V(S,O')))
is-committed(R, exact-paraphrase(O',O))
is-committed(R, relevant(V(S,O')))]
[Speech act: *inform*(R, audience, V(S,O'))]

This style sheet contains the pragmatic and situational information required to "understand" reported speech in newspaper articles. This is largely nondefeasible default knowledge as derived in Chapters 1, 3, and 4.

In a larger system, interfaces to underlying expert systems or other user programs could be connected in a similar fashion, namely as pragmatic information linked to the appropriate LCPs through semantic classes. If a belief maintenance system was the user of the lexicon, the *committment* slot could be filled with more explicit assumptions about beliefs, mutual beliefs, etc.

7.1.3 Semantic Class REPORTING–VERB

The semantic class reporting verb, another meta-lexical device, specifies inheritance from both, the LCP REP–VERB and the style sheet NEWS–REPORTING. The semantic

[2]The notation is illustrative, any implementation depends on the actual *interpretation* mechanism used.

class has to be specified in the lexical entry as a possible usage of the word.[3]

(3) **Semantic Class**

REPORTING–VERB

REP–VERB(A,B,C) & NEWS–REPORTING(X,Y,Z)
where A = X, B = Y, C = Z.

Semantic classes constitute an interface between the lexicon and pragmatic information and are used mainly by the parser and the common sense reasoning component. Semantic classes have to reflect the special format used by those components in order to express the compositional structure adequately, i.e. semantic classes have to provide hooks for user programs.

7.2 Reporting Verbs

Section 5.2.1 (p. 87ff) introduced some points that a lexical semantics for a semantic field must address, repeated here:

1. Define the semantic field, i.e. the semantic commonality that all members of the field share.
2. Define syntactic and other structural commonalities *typically* shared by members of the field.
3. Define the structure of the semantic field, i.e. give the relationships of the members towards other members.
4. Define a set of semantic dimensions along which the individual words are distinguished.
5. Define the individual words *as they deviate from the general pattern.*

[3]The representation of the evidential analysis of reported speech as an extension of the GL framework is inspired by an object-oriented programming style.

The general discussion of the evidential analysis of reported speech has addressed the first point mentioned there, namely the commonality that all members of that field share; the commonality is the *function* of these verbs to indicate *reliability* of the embedded statement by linking it to *circumstantial information* about the original utterance.

The second point, syntactic and other structural commonalities shared typically by members of the field, was captured by the defaults defined above in the semantic class REPORTING–VERB specified for each member of the field[4].

In this section I will explore the structure of the semantic field, reconsider the semantic dimensions introduced in Chapter 3, and finally define the individual words with emphasis on how they deviate from the general pattern and differ from each other.

7.2.1 Structure of the Field

The "hierarchies" that lexical definitions in conventional dictionaries form are entangled, shallow and often circular (cf. [Amsler, 1980]). Unfortunately, lexicographers do not consider semantic fields (or even parts of semantic fields); words are defined totally in isolation, no doubt the reason for some of the awkward two-item circular definitions and other shortcomings in conventional dictionaries.

Nevertheless, contrasting the structure of the "hierarchy" of genus terms in two dictionaries yields interesting insights into the implicit semantic dimensions involved.

Figure 7.2 shows the structure underlying the definitions in the Longman Dictionary of Contemporary English (LDOCE) [Procter, 1978]. The LDOCE is a learner's dictionary and thus keeps the lexical definitions simple, using only a core vocabulary of 2000 words for their definitions. This fact is clearly illustrated in Figure 7.2 by the central position of *state, make known,* and *declare*. LDOCE also focuses on *explaining* words, not *relating* them, resulting in fewer interconnections between words compared to the American Heritage Dictionary. Combined, these two factors explain why the reporting verbs considered form two separate hierarchies in LDOCE.

The roots of the two distinct hierarchies are *make known* and *state*. Several distinc-

[4]It is important to keep in mind here that I only consider the reporting verb meaning of the words. Thus, where other word meanings exist, this has to be clearly marked in a full entry.

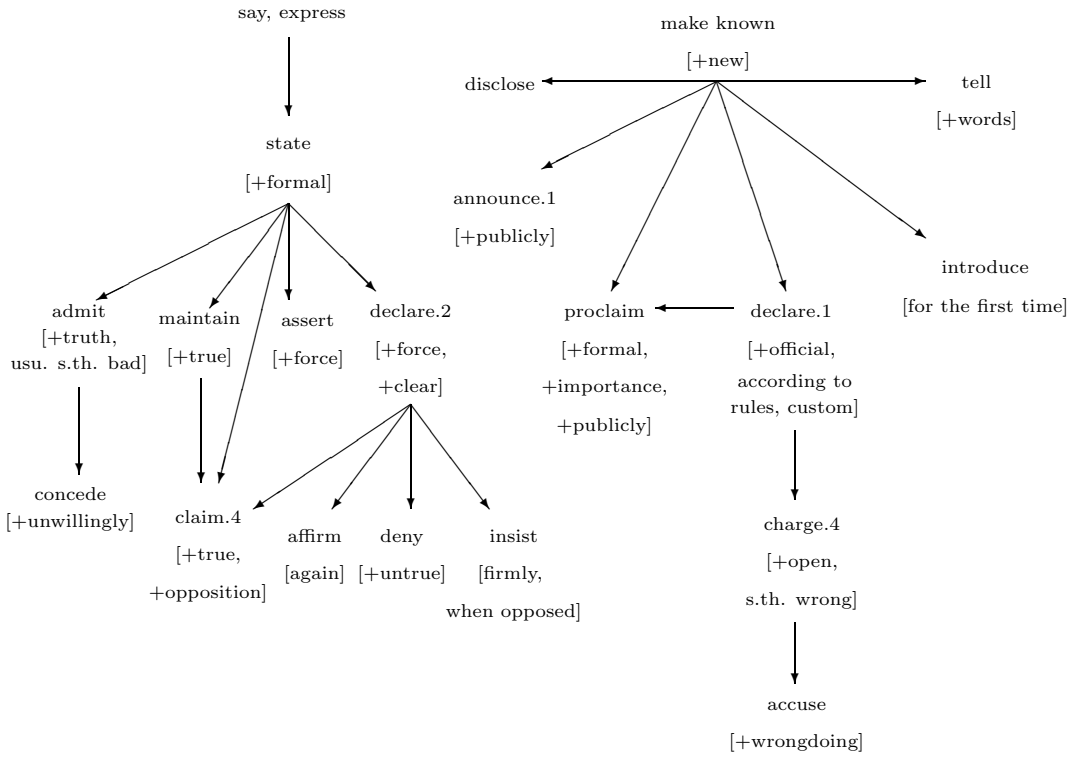


Figure 7.2: Hierarchy of genus terms in LDOCE.

tions can be construed: the hierarchy rooted in *make known* seems to focus on verbs that encode an *official* or *formal* aspect, whereas the verbs rooted in *state* focus on *truth*. But likewise, the *state* hierarchy contains those verbs that *presuppose* the proposition of the complement (or its negation) in the discourse situation, while the *make known* hierarchy implicitly encodes *novelty* of the embedded proposition. This reflects the different semantic dimensions mentioned briefly in Chapter 3.

Figure 7.3: Hierarchy of genus terms in the AHD.

Figure 7.3 describes the more connected structure underlying the American Heritage

Dictionary (AHD) [Berube, 1987][5]. *Make known* here is the root of the whole tangled hierarchy, focusing on the *information* aspect of reporting verbs. Two subhierarchies can be distinguished, rooted in *disclose* and in *state*. This is different from the distinction in LDOCE between new and presupposed information, here the distinction is made between *withheld* information and *new* information (where *new* here encompasses presupposed). Aside from this distinction, no further subhierarchies can be determined. The AHD does, however, have word clusters, groups of words that are defined in terms of each other and contrasted with each other. Word clusters are structures that one would expect to find in a semantic field oriented approach, where the basic meaning of the words is determined by the field and the individual word meanings are determined by how they differ from their closest neighbors. This approach is more appropriate for a dictionary that is intended to be used by native speakers, who will rarely be in need for a definition based on a restricted vocabulary, but will be able to appreciate close synonyms that convey finer aspects of the meaning. Thus, the definition

(4) **AHD:**
affirm: to declare or maintain to be true

conveys that to *affirm* can mean both putting forth the statement for the first time (declare) or repeatedly (maintain).

Another interesting observation is that *assert*, a term used often as a basic predicate (especially in truth oriented semantics) appears rather low on the AHD hierarchy and in the LDOCE hierarchy (where it also has no descendants within the described field), a reminder that terms used technically in the literature have to be studied very carefully to determine their true usage in everyday language.

7.2.2 Semantic Dimensions

The definition "hierarchies" of LDOCE and AHD show different distinguishing features, corresponding roughly to one or two of the semantic dimensions mentioned in Chap-

[5]Figure 7.3 does not contain *say*. This is because *say* is defined as: *1. To utter aloud. 2. To express in words. 3. To state; declare. 4. To recite. 5. To allege. 6. To indicate; show. 7. To estimate or suppose.* This forms more of a characterisation of the field than a definition of a word. I have omitted the many entries for *say*.

ter 3. In conventional dictionaries a semantic dimension can be expressed in two ways: either it is implicit in and inferred from the genus term or it is explicitly stated as differentia, modifications of the genus term. Usually we find both forms in the same lexical entry, the genus term implicitly defining the general, most salient features (determining the field) and the modification defining finer connotations. LDOCE encodes about the complement of the reporting verbs that it is assumed *true, untrue, new, bad,* or *wrong.* About the utterance situation the entries specify *formal, unwilling (source), public, official,* or *opposition.* The AHD encodes *true, untrue, fault, misdeed, offense, opposition,* and *positive* for the complement and *formal, official, authoritative, vehement* for the utterance situation.

The mechanism to define a semantic field through the default specifications of a semantic class together with the individual entry allows us to define a set of semantic features that are relevant for the entire field, thus making the semantic dimensions explicit in a computational lexicon. This structural way of delimiting semantic fields is somewhat different from previous studies such as [Miller and Johnson-Laird, 1976, p. 665ff].

Deriving principled set of semantic dimensions we observe that the reporting verb can specify four different elements of the reported speech sentence, namely the *physical characteristics of the original utterance*, the *original utterance situation*, the *attitude of the source towards the complement clause*[6], and the *strength of the complement.*

Physical Characteristics

The physical characteristics described by reporting verbs refer to the *voice quality* of the source at the time of utterance. The values for *voice quality* range over the entire spectrum of possible voice descriptions (high — *to cry*; low — *to wisper*; clear — *to enunciate*, unclear — *to mutter, to mumble*; high pitched — *to shriek*; low pitched —*to groan*; ...). The default for *voice quality* is *unmarked*, as in *to say.*

[6]This is why some of the reporting verbs discussed here would be called *attitudinals* in the literature.

Specifications of the Original Utterance

Further specification of the original utterance situation can be encoded in the reporting verb regarding *explicitness*, *formality*, and *audience*.

Explicitness is a scale ranging from *explicit* (*to explain, to elaborate*) to *implicit* (*to imply*) with all shades in between (*to hint, to suggest, to describe, ...*). An ordering of these entries according to increasing explicitness is not useful (and would in any case be subjective). A binary value *explicit, not explicit* is sufficient; the default is *explicit*, as is the case for *to say*.

Formality is a ternary feature with the unmarked default value *normal* and the two marked values *formal* (*to address*) and *informal* (*to blurt out*)[7].

The encoding of information about the audience in reporting verbs has three values, *public* (*to announce, to proclaim*), *private* (*to confide*), and the default value *unmarked*. The default assumption that the reporter witnessed the original utterance can always be overridden (*It is reported that X said ϕ, Apparently X said ϕ*) explicitly; this is therefore not part of the range of possible values for *audience*.

Specification of Attitudes toward the Complement Clause

The term *attitudes toward the complement clause* is used here much more restrictively than the term *attitude* is used in the general literature, compare for instance the use of *attitude* in TAMERLAN in Chapter 1.

Reporting verbs, here defined as utterance verbs and thus excluding cognitive verbs such as *to doubt*, encode attitudes of type *polarity, presupposition, speech act,* and *affectedness*. *Polarity* characterizes whether the speaker asserts the complement or its contrary, the values are *positive* (*to insist*) and *negative* (*to deny*). The default is *positive*.

Presupposition refers to the status of the reported clause in the context of the original utterance, the values are *new* (*to announce*) and *presupposed* (*to insist*) (cf. Chapter 5). The default value is *unmarked*.

[7]This is different from the register of the reporting verb itself, a feature that is called *familiar* or *slang* in dictionaries.

The value for the *speech act* dimension can be any of the generally assumed speech acts (cf. [Bach and Harnish, 1979]), the default is *inform*.

Affectedness has two values, *positive* (*to brag*) and *negative* (*to concede, to admit*) referring to the impact of the reporting clause on the source. The default value is *unmarked*.

Strength of the Complement

Strength of the complement refers to the reliability, certainty, or credibility of the complement as encoded in the reporting verb by the reporter. *To claim*, for example, has much less strength than *to state*, which is still lower in strength than *to announce*. This semantic dimension is only part of the evaluation of the reported speech as outlined in Section 3.3. In fact the semantic dimension does not even indicate the full impact the reporting verb has on the evaluation of the reliability, certainty, or credibility of the complement but only encodes whether the reporter leaves any room for doubt. The values are *high* strength (*to announce*) and *low* strength (*to claim*) with the default value *unmarked*.

Overview

Let me summarize the semantic dimensions outlined briefly in the previous paragraphs. The list of semantic dimensions is limited to those affecting the functional field of reporting verbs and even within this field no claims of completeness are intended. Semantic dimensions are in part subjective; different tasks warrant different granularity in the semantic dimensions. The dimensions given above provide a minimal partition of the functional field and are supported by the structures implicit in the definitional structure in LDOCE and the AHD.

The semantic dimensions of Figure 7.4 fall into two classes, which I will call *essential* and *optional* semantic dimensions. *Essential* semantic dimensions are those whose default value differs from *unmarked*, i.e. those dimensions that are by default specified by every reporting verb (*explicitness, polarity,* and *speech act*). It is no coincidence that the default values of these semantic dimensions are often seen as a pragmatic basis for

sem. dimension\values			default
voice quality	range : high, low, clear, ...		unmarked
explicitness	explicit	implied	explicit
formality	formal	informal	unmarked
audience	public	private	unmarked
polarity	positive	negative	positive
presupposition	new	presupposed	unmarked
speech act	range : request, question, ...		inform
affectedness	positive	negative	unmarked
strength	high	low	unmarked

Figure 7.4: Semantic dimensions of the functional field of reporting verbs.

successful communication: the Gricean maxims, for example directly specify two of the three default values ("be informative" specifies a default speech act for communication; "be relevant" subsumes that the salient information be explicit; cf. [Grice, 1967]). Positive polarity for the complement clause is generally assumed along with the default assumption that what is being said is also being asserted, i.e. has positive polarity.

Reporting verbs will generally only encode one *optional* semantic dimension, putting the specified optional dimension into *focus*.

Evaluation of Semantic Dimensions

The semantic dimensions can be used to evaluate the reliability implied by a particular reporting verb. I have pointed out in earlier chapters that an evaluation of reported speech is a subjective task. There are, however, some general criteria to the evaluation of the semantic dimensions that I will briefly outline here.

Chapter 6 discussed how discourse polarity items, reporting verbs that lexically specify for an opposition within the immediate context, by default lower the reliability assigned. Similar phenomena can be observed with different semantic dimensions.

Strength, by definition, indicates high or low reliability. *Negative affectedness*, by commonsense assumption, gives a high rating to the reliability, because we do not

assume that people say things that affect them negatively unless they are true. *Positive affectedness* might be interpreted not only as neutral, but as giving reason for caution. *Negative polarity*, because it violates a default assumption, emphasizes the possibility of an opposition (note that this is not the same as the implied opposition for *insist*) and thus slightly lowers the reliability rating. Utterances made in private have a lower reliability rating than utterances made to a public *audience*, but this default assumption can be easily overridden by context; information that one head of state receives from another head of state might very well be more reliable that what is publicly announced. And finally, *implied* statements are not as reliable as *explicit* statements, because they involve an extra interpretation step on the part of the journalist, a likely cause for error.

As these semantic dimensions combine in certain lexical entries (see next section), their individual contribution to the reliability rating are combined. *Announce* is a word that combines several strengthening semantic dimensions (audience: *public*, strength: *high*) and therefore receives a higher reliability rating than *maintain*, which has only one (audience: *public*). I have so far not encountered reporting verbs that combine a reliability strengthening semantic dimension with a reliability lowering one.

The preceding discussion has to be put into perspective: I am not suggesting here that assessing the reliability of reported speech is as easy as counting the number of strengthening or weakening semantic dimensions in the reporting verb. The reporting verb builds an evaluative environment for the reported material only in composition with the source description. Thus by itself the semantic contribution of the reporting verb is rather small. Moreover, the utterance context can highlight certain semantic dimensions and alter their contribution to the evaluative environment, as the example of the summit meeting of heads of state indicates. Nevertheless it appears that semantic dimensions guide our assessment of the reliability of reported speech. A rating of the semantic dimensions would require careful studies on different corpora which go beyond the scope of this thesis.

7.3 Lexical Entries for Reporting Verbs

This section introduces lexical entries for a set of reporting verbs in Generative Lexicon notation. The list of reporting verbs is not complete, rather some interesting observa-

tions are detailed for some frequent reporting verbs drawing on corpus data from the Wall Street Journal corpus and represented in the GL formalism.

7.3.1 *Say*

Say is the most unmarked and the most frequent reporting verb. The American Heritage Dictionary Pocket Edition (AHD) defines *say*:

> **say** ...1. To utter aloud. 2. To express in words. 3. To state; declare. 4. To recite. 5. To allege. 6. To indicate; show: *The clock said noon.* 7. To estimate or suppose.

According to this definition *saying* can be a whole range of things. The central word senses for reporting verbs are sense (2) and (3). Some example sentences are:

(5) (a) "The message to the board of education out of all this is we've got to take a serious look at how we're doing our curriculum and our testing policies in this state," said the talk-show host.

(b) The stocks of banking concerns based in Massachusetts weren't helped much by the announcement, traders said, because many of those concerns have financial problems tied to their real-estate loan portfolios, making them unattractive takeover targets.

(c) A bank spokeswoman also declined to comment on any merger-related matters, but said the company decided to drop its opposition to the interstate banking legislation because "prevailing sentiment is in favor of passage."

(d) SHAREDATA Inc. said it will amend a registration statement filed with the Securities and Exchange Commission to delete a plan to sell 500,000 newly issued common shares.

The word sense (1) in AHD for *say*, where *say* is synonymous with *utter* cannot occur in the particular syntactic construction that I have selected here. When a full clausal complement is realized, only the reporting sense of *say* is applicable. The reporting verb covers not only senses (2) and (3), but also (5) and (7); these senses have to be clearly marked in the immediate context in order to be recoverable. I will therefore subsume these stronger senses under the reporting verb sense and rely that the analysis component will recover these meanings from the context.

In the reporting verb sense for *say* there is a default assumption that what was *said* was also *asserted* (sense 3 in the AHD). The lexical entry for *say* is therefore given as[8]:

say(T,A,B)
[Form: REPORTING–VERB]
[Telic: default: say(A, true(B))]

Because *say* is the most unmarked and the most frequent reporting verb, one would expect to find other reporting verbs defined using *say* as genus term. Yet we find that this is almost never the case because in common use *say* contrasts with *mean*, as in *He just said that . . . , he didn't mean it.* While this sense of *say* is only licensed in contrastive environments (as the example above or an unusually stressed *say*, for instance *So he* SAYS), dictionary entries still opt to use more precise and contentfull genus terms.

7.3.2 *State*

To state something means asserting the truth of the stated matter, as evidenced by the usage of (6b) *incorrectly stated*. The statement can take the form of an utterance or it can be made in writing. Indeed, *state* is frequently used together with a modifying prepositional phrase indicating a kind of document, as in (6a), (b), and (e). In a metonymic extension the document containing the statement can also appear in subject position, as in (6f).

(6) (a) "The USIA publicly and officially stated in the litigation that all persons are allowed access to the materials . . . " Mr. McCormick noted.
(b) In yesterday's edition, it was incorrectly stated that SunAmerica was being divested.

[8]The lexical entries presented here are not general entries that one would expect to find in a Generative Lexicon but are once again geared towards the topic of this thesis. In a general GL implementation an entry for *say* might be given with the literal sense (1), synonymous with *utter*. The argument structure will indicate that there are two syntactic patterns associated with *say*, namely an NP complement or a sentential complement (with optional *that*). A meta-lexical rule would then allow coercion from the literal sense to the reporting verb sense generalizing over the fact that all utterance verbs that can take a sentential complement can be used as reporting verbs in our definition. A yet more general definition would allow utterance verbs with NP complements as reporting verbs if the NP was of appropriate type. The entries here are therefore much simplified by giving the reporting verb reading explicitly in the entry rather than indirectly through an LCP mapping the entry to the reporting verb reading.

(c) The police report, filed on June 4, quotes Mr. Wynn saying that he "bought a baby for three dollars" from a woman who "stated that she have a baby she will kill if she don't get three dollars."
(d) As Sen. Robert Dole (R., Kan.) aptly stated on the Senate floor, attorneys will not be able to build careers out of lawsuits against public accommodations.
(e) He pointed out that in a communique issued on Sept. 23 in Washington, the G-7 countries stated that the dollar's strength conflicted with economic fundamentals, meaning the trade imbalances between the U.S. and its major trading partners.
(f) VICTOR SPERANDEO recommended selling short the stock of Travelers Corp., as stated in yesterday's Investment Dartboard column.

An interesting syntactic alternation to the pattern $<$ *source* $>$ *stated* (*that*) ϕ is *As* $<$ *source* $>$ *stated* (...), ϕ as in (6d) and (f).

The AHD defines *state* as:

state To set forth in words; declare.

Thus *state* is synonymous with the reporting verb sense of *say* (with a slightly stronger feeling that the complement has been claimed to be true):

state(T,A,B)
[Form: REPORTING–VERB]
[Telic: say(A, true(B))]

State cannot stand with direct quotes; the argument structure in the LCP for *state* has to reflect that fact.

7.3.3 *Assert*

Assert has come to mean *to establish the truth of* in the linguistic literature, where it is usually used as the strongest term to express truth. Yet usage in the Wall Street Journal corpus does not seem to distinguish between *say, state,* and *assert*.

(7) (a) He also asserted that exact questions weren't replicated.
(b) He asserts that government has done an even worse job of controlling its health bill than business.

(c) But the Wayne State group asserts its finding differs from the earlier ones.

(d) "The agricultural sector cannot progress under the current level of funding," the report asserted.

(e) Although officials assert that the environment is cleaner today than it once was, such claims seem dubious in light of some grim statistics.

(7a)–(c) could be rephrased using *say*, (7d) using *state*, and (7e) could be rephrased using *claim* or *insist*, both verbs encoding a higher degree of uncertainty than is usually associated with *assert*. It appears then that the Wall Street Journal uses *assert* as a synonym to *say* and *state*. This is in accordance with the dictionary definition found in the AHD:

assert 1. To state positively; affirm. 2. To defend or maintain.

Assert can only be used in reporting situtations, be it the report of the language of others or the report of the speaker's own intentions. It is this last case that has found most attention in the speech act literature; I will ignore the complications arising from this use here.

assert(T,A,B)
[Form: REPORTING-VERB]
[Telic: say(A, true(B))]

7.3.4 *Affirm*

Affirm does not occur very frequently in its reporting verb sense, more frequent are the adjective *affirmative* and the noun affirmation. Within the reprorting verb sense it appears more frequently with a nominal complement, often in the sense of *to confirm*, as in (8a), (c), and (d).

(8) (a) A Louisiana judge affirmed a decision by state regulators to prevent the utility from recovering $1.4 billion in costs associated with its River Bend nuclear plant through rate increases.

(b) "In Mexico, we now affirm that drug-trafficking is a threat to national sovereignty because, as in other nations, it corrupts whatever it touches," Salinas says.

(c) UNITED TECHNOLOGIES Corp.'s double-A-minus senior debt rating, single-A-plus subordinated debt rating and single-A-1-plus commercial paper rating were affirmed by Standard & Poor's Corp.
(d) Standard & Poor's Corp. affirmed its rating on the company.

Common to all examples is that *affirm* specifies an optional semantic dimension, namely that it *presupposes* the complement. This is also implied in the definition in the AHD:

AHD:

affirm 1. To declare or maintain to be true. 2. To ratify or confirm.

The lexical entry is given as:

affirm(T,A,B)
[Form: REPORTING-VERB]
[Telic: say(A, true(B))] & presupposed(B)]

Very similar to the usage of *affirm* is *reaffirm*, which occurs slightly more frequently in the sample chosen:

(9) (a) The judge also reaffirmed the five-year rate-increase plan previously approved by the Louisiana Public Service Commission.
(b) The country's hard-line Communist leadership reaffirmed the nation's commitment to socialism, but said demands for the types of changes sweeping much of Eastern Europe "are open to discussions."
(c) The paintings reaffirm in the most direct fashion that Velazquez is perhaps the greatest optical realist in Western art.
(d) Salomon Brothers reaffirmed its buy recommendation on the stock.

The difference in meaning between *affirm* and *reaffirm* is that *reaffirm* specifies that the complement is not only generally presupposed, but has been *affirmed* before; a regular use of the prefix *re*. The entry is given as:

reaffirm(T,A,B)
[Form: REPORTING-VERB]
[Telic: repeat(affirm(A,B))]

7.3.5 *Announce*

Chapter 5 introduced a preliminary analysis of *announce*, specifying the optional semantic dimension *presupposition* as *new*. Additionally, *announce* is a speech act and requires that the source has some legitimation to announce the proposition of the complement. Probably due to this last fact *announce* also specifies *high* as the value for the semantic dimension *strength*. Finally, an *announcement* has to be made *publicly* to be properly called an announcement (or intended for further publication).

(10) (a) U.S. News has yet to announce its 1990 ad rates.
(b) However, five other countries — China, Thailand, India, Brazil and Mexico — will remain on that so-called priority watch list as a result of an interim review, U.S. Trade Representative Carla Hills announced.
(c) The "one-yen" controversy first came to a head last week when the city of Hiroshima announced that Fujitsu won a contract to design a computer system to map its waterworks.
(d) When Warren Winiarski, proprietor of Stag's Leap Wine Cellars in Napa Valley, announced a $75 price tag for his 1985 Cask 23 Cabernet this fall, few wine shops and restaurants around the country balked.

The revised entry for announce is given as:

announce(T,A,B)
[Form: REPORTING–VERB & public(T)]
[Telic: say(A, true(B)) & new(B)]
[Constitutive: SPEECH-ACT: announce & STRENGTH: high]
[Agentive: legitimation(A,T)]

7.3.6 *Claim*

Claim is an interesting verb because it combines three very different word senses. The AHD gives as a definition:

claim 1. To demand as one's due. 2. To state to be true. 3. To call for; require.

I would draw the distinction slightly differently, distinguishing the reporting verb sense (sense (2) in the AHD), the sense of demanding or requesting in front of some authority, and the sense of having successfully claimed, i.e. having taken. This distinction

can be oberved in the Wall Street Journal data, where (11a) and (d) correspond to the reporting verb sense, (b) and (c) correspond to the sense of demanding or requesting in front of authority, and (11e) corresponds to the successful claim sense. I will disregard the non-reporting verb senses in the entry for claim.

(11) (a) He says he had Candlestick built because the Giants claimed they needed 10,000 parking spaces.
(b) He claimed losses totaling $42,455 — and the IRS denied them all.
(c) Mr. Dworkin responded in June by filing suit in federal court in Los Angeles in which he claimed wrongful dismissal.
(d) That represents a very thin "excess" return, certainly far less than what most fundamental stock pickers claim to seek as their performance objective.
(e) The storm has claimed the lives of some 24 people in the U.S. and has left tens of thousands without phones or electricity.

Claim in its reporting verb sense also specifies the semantic dimension *strength*, indicating *low strength*.

Claim falls outside the standard syntactic patterns for reporting verbs in that it accepts an infinitival complement; this fact would have to be represented in the appropriate LCP in a fully developed lexicon. Here it suffices to give as the entry for claim:

claim(T,A,B)
[Form: REPORTING–VERB]
[Telic: say(A, true(B))]
[Constitutive: STRENGTH: low]

7.3.7 *Insist*

Chapter 6 introduced a detailed analysis of *insist*, stating that inherent in the meaning of *insist* is an opposition between the complement and some accessible proposition in the context, thus also specifying the semantic dimension *presupposition*. The revised entry is

insist(T,A,B)
[Form: REPORTING–VERB]
[Telic: say(A,true(B)) & presupposed(B) & opposed(B,ψ)]
[Const: MANNER: vehement]

7.3.8 *Deny*

Deny, in contrast to *insist*, does not imply an opposition to a contextually availabe statement but it actively creates one. This means that the complement is *presupposed* and the *polarity* is *negative*. Since the presupposed statement is often mentioned just prior in the text, a summarizing event nominal often takes the place of the complement, as in (12d).

(12) (a) Viacom denies it's using pressure tactics.
(b) The U.S. wants the removal of what it perceives as barriers to investment; Japan denies there are real barriers.
(c) I had sought, in my suit, the right to print Voice material, which had been denied me, and I had sought a right to receive the information, arguing in effect that a right to print government information isn't very helpful if I have no right to get the information.
(d) A Fed spokesman denied Mr. LaFalce's statement.

The AHD definition for *deny* is

deny 1. To declare untrue; contradict.

The entry can therefore be given as

deny(T,A,B)
[Form: REPORTING–VERB & POLARITY: negative]
[Telic: say(A, not(B)) & presupposed(B)]

and is identical to the synonym *contradict*:

contradict(T,A,B)
[Form: REPORTING–VERB & POLARITY: negative]
[Telic: say(A, not(B)) & presupposed(B)]

7.3.9 *Maintain*

Maintain is another verb that occurs with very different meanings. The AHD lists the reporting verb sense only as 6th (and last) meaning.

maintain 6. To assert or declare.

However, examples of the reporting verb sense are common in the Wall Street Journal corpus and imply that the complement is *presupposed* and uttered *publicly*:

(13) (a) Several traders maintained that the Merc's 12-point circuit-breaker aggravated the market slide Oct. 13 by directing additional selling pressure to the floor of the New York Stock Exchange.
(b) Many of the letters maintain that investor confidence has been so shaken by the 1987 stock market crash — and the markets already so stacked against the little guy — that any decrease in information on insider-trading patterns might prompt individuals to get out of stocks altogether.
(c) But Judge Keenan said that privilege is meant to protect private utterances — not litigation papers filed with foreign governments, as Mrs. Marcos's attorneys maintained.

The entry is therefore represented as

maintain(T,A,B)
[Form: REPORTING–VERB & public(T)]
[Telic: say(A, true(B)) & presupposed(B)]

7.3.10 *Admit*

Admit is a verb that specifies *negative affectedness* by the complement for the source, being defined in the AHD as

admit 4. To acknowledge; confess. 5. To concede.

In the corpus data we find that the prevalent word sense for *admit* is (5), where the complement is also *presupposed*. The *strength* for *admit* (as for most reporting verbs that affect the source negatively) is *high*.

(14) (a) Mrs. Yeargin admits she made a big mistake but insists her motives were correct.
(b) He admits, though, it isn't one of Campbell Soup's better products in terms of recyclability.
(c) Vicar Marshall admits to mixed feelings about this issue, since he is both a vicar and an active bell-ringer himself.

(d) Without admitting or denying wrongdoing, they consented to findings that they failed to return funds owed to customers in connection with a limited-partnership offering.

(e) Glasnost may be partly responsible, but Soviet Foreign Minister Eduard Shevardnadze last year admitted, "The exaggerated ideological approach undermined tolerance intrinsic to UNESCO."

The entry is therefore represented as

admit(T,A,B)
[Form: REPORTING–VERB]
[Telic: say(A, true(B)) & presupposed(B) & AFFECTEDNESS: negative]
[Constitutive: STRENGTH: high]

7.4 Source NPs

This section presents an analysis of some frequent source descriptors we find in the Wall Street Journal Corpus paying special attention to nouns specifying expertise or authority in some fashion.

The influence of the complex source noun phrase on the trustworthiness of the attributed utterance is very subtle and can not be derived solely from the lexical semantics of the words involved. The objective here is to find the contribution that certain lexical items and their composition add to this process. The approach is to look at corpus data and to extract the patterns that we find. This step is done automatically, using a partial parser for noun phrases based on the tags in the tagged corpus to extract the "noun phrases" that contain the seven nouns. Careful manual analysis then determines the different paradigms of the source NPs.

Based on the findings I present lexical entries for the seven nouns, defining their lexical contribution to the evaluation of reported speech.[9] Of particular interest are constraints on the noun phrases containing the nouns, especially constraints on the realization of possible arguments. The notion of argument here is somewhat wider than in the traditional linguistic literature. I call a *semantic* argument those words or concepts that have to be cospecified in the text for a noun to make sense. *Semantic*

[9]Note that I do not claim completeness for the lexical entries developed in this section.

arguments do not have to occur in the surface structure but have to be inferrable. For example *spokesman*, as I will argue in a later section, requires an employer argument to be inferable.

7.4.1 *Official*

One notorious source in newspaper articles is the "official" or preferably a whole group of "officials". The "official" is usually implicitly trustworthy, even when his or her relationship to the subject matter of the article is not at all elucidated (compare the discussion in Chapter 5). This trustworthiness is already apparent in the dictionary definitions and therefore, I claim, lexicalized.

LDOCE defines *official* as:

LDOCE:

> **official n** a person who works in government, esp. one whose job is less important than that of a government OFFICER — see OFFICER (usage)
> **official adj** of or about a position of trust, power, and responsability: *an official position/an official occasion/an official manner of speaking.* — opposite: **unofficial**; compare officious

The AHD definition of *official* is

> **official n** 1. One who holds an office or position.
>
> **office** 3. A position of authority given to a person, as in a government or other organization. 4. A branch of the U.S. government ranking just below a department.

The main meaning of the word *official* when used as a source description in reported speech is to convey that the source holds a position of authority or trust and that the reporter wants the reader to accept this for the sake of the article even though he or she might not give any more justification for that trustworthiness than the description *official*. Note that authority and trust are imparted on the source by his/her employer. Trustworthiness with respect to the newsstory arises out of the source's closeness to information that is relevant, as discussed in Chapter 5. Thus, even though the reader

may personally distrust the source, the reader should still acknowledge the source's inherent authority (compare Text 1 on Page 51).

Official takes one optional semantic argument, namely the institution, in which he or she is employed. There is also an implicit argument, namely the particular position that the "official" holds. It is this position, that has authority, but its nature is not revealed by the use of the word *official*.

The syntactic paradgims in which *official* occurs are simple. Modifying information can either occur prenominally or as a directly attached prepositional phrase.

1. a university official
2. a Kemper official
3. several officials from Jenney
4. an official of the Ministry of Health and Welfare

The LCPs for *official* and the argument to *official*, y, are:

(determiner) (rank) reference-to(y) *official*

or

(determiner) *official* ($from \mid of \mid at \mid with \mid \ldots$ reference-to(y))

While the name of the employer is the most frequent surface realization of the argument, this realization is not always direct. We find besides the name also the type or category of y, the location of the offices of y, the region of distribution of the goods or services provided by y, or other attributes of y:

1. the Treasury official
2. White House officials
3. Mexican officials
4. the ruling party officials

5. A federation official
6. Communist officials
7. Washington officials
8. federal officials in Washington
9. The regional officials

We find that the range of possible references to the argument must be lexically specified when we compare the data for *official* with the data for *spokesman* in Section **??**. There we do not find the same range of referring expressions. Indeed, many of the constructions are not possible:

1. the Treasury spokesman
2. ? a White House spokesman
3. * a Mexican spokesman[10]
4. the ruling party's spokesman
5. a federation spokesman
6. * the/a Communist spokesman
7. * a Washington spokesman
8. * the federal spokesman in Washington
9. * the regional spokesman

Let us formulate the semantic constraints pertaining to the reference to the argument of *official* as constraints on the LCPs, and add to the LCP above the following elaboration:

reference-to(y) $\longrightarrow$ (region1) (region2) (type | goods | additional attributes) (name)

This is the point where lexical semantics and knowledge representation overlap: the different references to the argument of *official*, the employer, define a template for *employer*. We describe in the next section, that the different kinds of references contribute additional information that is relevant to the text. The chain of reference builds a profile of the employer. We have already used the slot names of the *employer template* in our characterization of the constraints to the LCP for *official*. Here is the template itself:

[10] *A Mexican spokesman* will usually denote a spokesman who only happens to be a Mexican, and not a spokesman for Mexico.

Employer template for *official*
NAME:
TYPE: structure/category of business
GOODS: service or product rendered
REGION1: location of office
REGION2: region of distribution of goods or services
ADDITIONAL ATTRIBUTES: for instance time(-interval) when description was valid, current person in top position etc.

A GL entry for *official* then takes the following form:

official(*x*, y)
[Form: *human(*x*), organization(y),*
*∃position(z):in(z,y) ∧ has-authority(z) ∧ hold(*x*, z)*]
[Telic: *work-for(*x*, y), trust(y, *x*)*]
[Constitutive: *individual(*x*)*]
[Agentive: ↑ *y*]

7.4.2 *Spokesman, Spokeswoman*

The "spokesman" or "spokeswoman" is another frequent source of information. Unlike an "official", a "spokesperson" is usually clearly specified — the individual could usually be identified even if the name was not given in the article. Moreover, *spokesman* and *spokeswoman* describe concrete positions with a clear set of duties. This is reflected in the dictionaries through simple, single entries.

The dictionaries define *spokesman* as follows:

LDOCE

> **spokesman** *fem* **spokeswoman** a person chosen to speak and represent the opinions of others officially.

AHD

> **spokesman** One who speaks on behalf of another or others. — **spokeswoman**

Important to the role of a "spokesman/spokeswoman" as a source of information is that:

a) a spokesman/spokeswoman is "chosen"

b) a spokesman/spokeswoman "represents" somebody else "officially"

Both (a) and (b) imply that the "spokesman" "spokeswoman" has reliable information on the group or organization that he or she represents and that he or she has the trust of that group or organization.

Spokesman, spokeswoman, like *official*, has one argument, namely the group or organization that he or she represents. In contrast to *official*, however, it appears more often without the argument realized in the same NP in the Wall Street Journal Corpus.

The argument can appear both in prenominal position or in a directly attached PP. The variety of expressions referring to the argument (i.e. the employer) is much more restricted than for *official*: we find only names or direct descriptions of the group or organization (i.e. the type, *company*, or the purpose, *the retail jeweler*). While locatives do occur, they do not refer metonymically to the employer argument even in cases of subsequent reference. "Spokesmen/spokeswomen" are also more often introduced by name, than officials. Additional information about the group or organization they represent is usually added as an apposition or relative clause.

1. a Lincoln spokesman
2. a White House spokesman
3. White House spokesman Marlin Fitzwater
4. a government spokesman

5. a spokeswoman for Miller Brewing Co.
6. a spokeswoman for Millicom, a telecommunications company
7. David Bell, a spokesman for the airline

The LCPs for spokesman/spokeswoman consequently are:

(determiner) (representee) spokesman/spokeswoman (name)
or
(name ",")(determiner) spokesman/spokeswoman (preposition representee)
where
representee $\longrightarrow$ (NAME | TYPE)

where NAME and TYPE are again slotnames in the knowledge representation frame for the entity represented by the "spokesperson". Thus the kind of information that can be conveyed indirectly about the employer of the spokesperson by way of modifying the noun phrase containing the word *spokesman/spokeswoman* is very limited:

Employer template for *spokesman/spokeswoman*

NAME**:** name of the group or organization

TYPE**:** category of the goup or organization

The GL entry for *spokesman* (identical to that of *spokeswoman* is given as:

spokesman(*x*, y)
[Form: *human(*x*), human(y), default: organization(y)*]
[Telic: *speak-for(*x*, y), represent(*x*, y, press and/or public)*]
[Constitutive: *individual(*x*)*]
[Agentive: ↑ y]

7.4.3 *Analyst*

The word *analyst* occurs 1203 times in the WSJC. It is thus even more frequent than *official*. This may indicate a certain bias of the WSJ, with financial and market analysts playing a greater role than in other newspapers. Unfortunately, the dictionary

definitions described here do not directly support the use of *analyst* as encountered in the WSJ.

Dictionary definitions for analyst are:

LDOCE

> **analyst** a person who makes an ANALYSIS, esp. of chemical materials: *a chemical analyst*
> **analysis** 1. a separation of a substance into parts: *The analysis of the food showed the presence of poison* —compare SYNTHESIS 2. the results of such a separation, esp. as a list 3. an examination of something together with thoughts and judgements about it

AHD

> **analysis** 1. The separation of a whole into constituents with a view to its examination and interpretation. 2. A statement of the results of such a study. . . . — **analyst** — **analytic**

Analyst implies expertise. This is only implicit in the dictionary definitions but it is the basis for the term *analyst*: an "analyst" is not somebody who happens to have separated something into its parts, but an "analyst" is a person who has studied a field extensively, giving him or her the insight necessary to analyse situations and also does these analyses on a regular basis.

We find that *analyst* appears in the corpus frequently with at least one argument, namely the kind of data analyzed. This argument, when realized, occurs prenominally:

1. some market analysts
2. election analysts
3. airline analysts
4. currency analysts

There is, however, another kind of (postnominal) modification that occurs very frequently in the corpus. This is the specification of the employer of the *analyst*. We find the folowing forms:

1. an analyst at First Manhattan Co.

2. analyst at the brokerage Cholet-Duopnt & Cie.
3. an analyst with Phoenix Capital Corp

The employer can also be specified prenominally and both, the specialty and the employer can be specified at the same time. Additionally, we find locative PPs, but not used metonymically (but see *European analysts, U.S. analysts...*) in the data analyzed so far.

1. senior quantitative analyst at Merrill Lynch & Co
2. PaineWebber analyst Thomas Doerflinger
3. Merrill Lynch food analyst William Maguire Lawrence Eckenfelder,
4. a securities industry analyst at Prudential-Bache Securities Inc.
5. a government analyst
6. a Commerce Department analyst

It seems that to specify the employer is often sufficient to indicate the field of expertise of the analyst. But because there is no true logical correlation (a bank might employ a food analyst etc) I regard the employer as a second argument to *analyst*.

The LCPs for *analyst* are:

(determiner) (rank) (employer) (field of specialty) *analyst*

or

(determiner) (rank) (field of specialty) *analyst* (employer)

Here, the different realizations we find for the *employer* are very limited. In most cases we find the name of the company. Only in very few cases can we find a common noun, such as *government, industry,* or *private*.

The restrictional template for the employer as argument to *analyst* is therefore the same as that for *spokesman*[11]

[11]Note, that these restrictional templates are a generalization over the data. While there is a slight difference between the realizations of the employers as argument to *spokesman* and the employers as argument to *analyst*, they are indistinguishable on our account. A more precise description of the data might associate usage counts on different corpora with the different template slots. But to my mind, this would benefit neither analysis nor generation.

Template for employer argument to *analyst*

NAME**:** name of the group/organization

TYPE**:** category of the goup/organization

The GL entry for *analyst* is therefore:

analyst(*x*, y, z)
[Form: *+ profession, employee(*x*, y), expert(*x*, z)*]
[Telic: *analyze(*x*, z)*]
[Constitutive: *individual*]
[Agentive: ↑ *y*]

Chapter 8

Representing Reported Speech

The evidential analysis proposed for reported speech in newspaper articles assumes a two step process. First, the linguistic characteristics have to be processed, resulting in the representation of evidential scope. Then a reasoning process has to evaluate the evidential scope and assign a reliability to the reported material. I argue that the separation of the two steps is essential, because the final evaluation task is subjective, requiring world knowledge and personal belief assumptions that change from reader to reader. And the same reader might very well interpret the same instance of reported speech differently in different contexts.

This chapter extends this two step model to text analysis in general. *Minimal Text Representation* (MTR) is introduced, a representation scheme to adequately represent the linguistic structure of a text. Applying the notions of MTR to the analysis of reported speech I will show that a lazy evaluation strategy can be implemented for the representation of reported speech. This representation also creates a notion of local context that is particularly useful for retrieval even under partial evaluation.

8.1 Minimal Text Representation

Minimal Text Representation (MTR) aims at representing written texts in a way that preserves all information provided in the text in a form suitable for further processing by

different NLP *user modules*. Such user modules could be in form of belief representation or maintenance systems, summarizing systems, or information extraction systems (such as, for example, TIPSTER).

The idea behind MTR is to provide a multi-purpose representation of the linguistic aspects of a text that can be used and re–used by many different reasoning and interpretation systems. In the traditional sense MTR does not represent the meaning of the text, but rather a preprocessing stage. Figure 8.1 illustrates the point.

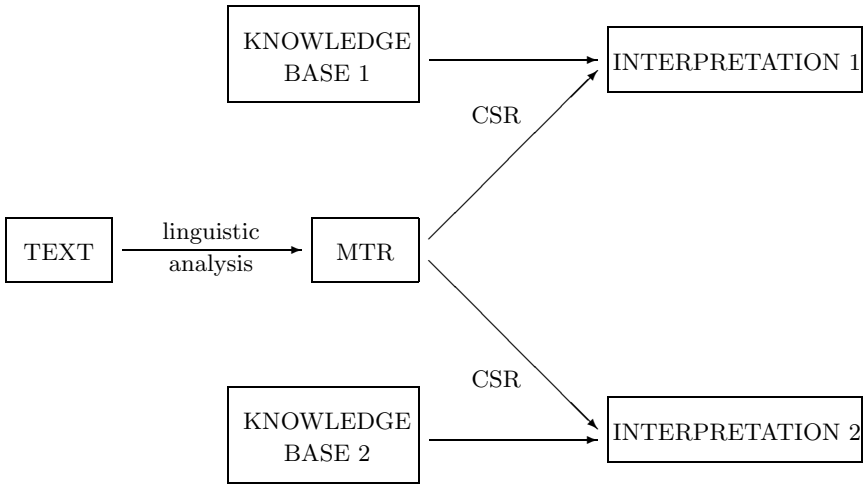


Figure 8.1: MTR represents a preprocessed version of the text for later interpretation under different knowledge bases and belief systems.

MTR is designed to mirror the information given in the text closely, both in content and in structure. Preserving the original text structure is important for interpretations that are sensitive to style. However, MTR will interpret and make explicit implications in the language of the text; for instance the coherence constraint imposed on the arguments of *announce* (discussed in Section 5.4) will be represented. Thus, MTR provides inferences that are drawn from the text but it draws its inferences solely from the lexical compositional semantics of the words involved, keeping the amount of *commonsense* inferencing to a minimum, thereby forcing the user modules to do additional inferencing suitable to their task. This is indeed a positive feature, allowing the same text representation to be used to extract different information. Reusability has only recently become a concern in the computational linguistics community, mainly motivated by the need to share tools and data for the analysis and exploitation of very large on-line corpora.

MTR can be seen as an important step on the way to producing more reusable resources by being *usable* in different contexts and by very different systems.

MTR is a representational scheme, not a fully developed representation language. It augments traditional representation techniques and is intended to work in cooperation with existing representation languages.

The goal of MTR to represent the text faithfully and in an unbiased fashion is of course an idealization. There is no unbiased representation — in fact it is questionable whether there can be any useful representation that does not have a purpose or task "built in". MTR attempts to introduce as little bias as possible into the representation by representing facts and inferences that project up from the lexical semantics and the compositional behavior only. The task of MTR is at this stage to make explicit certain important characteristics of reported speech in accordance with an evidential analysis. Future work will extend the notions of MTR to a treatment of modals, evidentials in general, and negation.

One key idea behind MTR is to disentangle the representation of different aspects of a text. As we have seen in Chapter 3 when discussing the representation of intensional contexts in Discourse Representation Theory, a comprehensive representation is very complex and unwieldy. In addition, comprehensive representations often tend to be brittle; again considering Discourse Representation Theory, it is not clear how a partial representation could be achieved that represents the temporal relationship between clauses whose embedding relationship cannot be determined, for instance. The DRT model is of course not concerned with such cases; however, an information retrieval use of DRT will not be feasible in the near future exactly because DRT assumes a comprehensive analysis.

MTR provides several independent representation devices[1] based on an earlier model described in [Bergler and Pustejovsky, 1990]. The three devices are:

Coherence Structure: The coherence structure captures the relations between clauses and the linear structure of the original text.

Trace: A trace captures the sequence of temporal activities which can be put into a

[1]These devices are independent in representation but interconnected with multiple links.

partial order on the narrative time line. A trace is constructed from *trace segments* in a bottom-up fashion using *trace composition* and *trace unification*.

Profile: A profile contains a list of all properties the text asserts or implies about a particular discourse entity. Distinct discourse entities have separate profiles.

To adapt this model to handle reported speech, we have to consider the evidential analysis in this light. We have seen that an individual instance of reported speech consists of three structural parts with possibly additional information: the source, the reporting verb, and the reported material of the complement clause. In Chapter 3 evidential scope was defined to express the evidential analysis of reported speech, capturing the context split in so called context variables. The interpretation of these context variables was judged to be dependent on the *world view* of the evaluating agent as well as on the purpose for which the extracted information would serve.

MTR is one way to implement an evidential analysis, focusing on the role of the source in the evaluation of reported speech. In the following section I will briefly describe two MTR devices mentioned above, focusing mainly on the profile structure, and how it achieves a representation of reported speech. I will not discuss coherence structure in any detail, assuming a conventional model as outlined in [Hobbs, 1982]. The MTR devices outlined here are indeed overly simplified. I will discuss shortcomings and refinements when presenting some real text analyses in Chapter 9.

8.2 Trace

A *trace* represents a partial temporal ordering of the events of a text. The form of a trace is that of a graph. Individual trace segments are partial time lines, on which text events are ordered. Only explicit temporal orderings are represented on a trace. If a text does not specify a temporal ordering between two events explicitly, the events will be placed on two different trace segments. Trace segments intersect at the closest point in the trace with respect to which they stand in a known temoral relation.

Traces are constructed bottom up, starting at the clause level. Lexical semantics, tense information and explicit temporal modification result in the ordering of individual

trace segments. This is a very difficult problem and has been discussed in the literature [Hinrichs, 1986, Reyle, 1986, Webber, 1988, Moens and Steedman, 1988]. This thesis does not contribute to this discussion but acknowledges that temporal ordering is necessary for any representation system; a partial ordering that allows unresolved temporal relationships to be represented without committing to any ordering seems best suited to the goals of MTR.

As an illustration of a very simple trace consider Figure 8.2. For more detail on how trace segments can be manipulated to partially resolve temporal relationships see [Bergler and Pustejovsky, 1990].

Who's News: Pacific Enterprises Chooses Ukropina As Chief Executive

Pacific Enterprises named its president, James R. Ukropina, as chairman and chief executive, succeeding Paul A. Miller and ending a century of family leadership at the utility holding company.

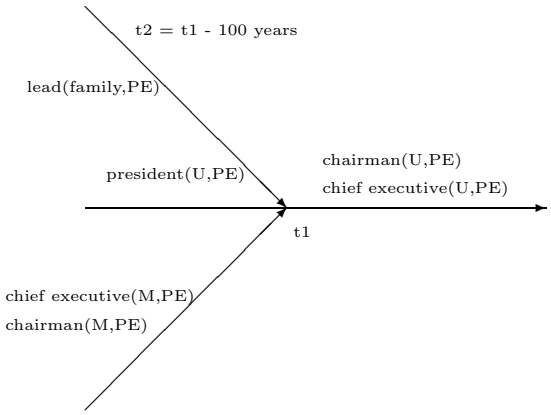


Figure 8.2: Simple trace fragment.

8.3 Profiles

We have seen that the source is a major part of the evaluative environment of primary information in newspapers. This importance of the source is mirrored in MTR in the construct of the *profile*.

Profiles are a collection of information that has to do with one specific discourse entity: attributes and features of that entity, such as being blond or being a CEO; expressed thoughts and beliefs attributed to the entity, such as being concerned about the environment or believing that the recession will end soon; and generally all utterances made by that entity, such as announcing a joint venture or saying "Oh, no!".

Profiles differ from belief contexts in that they contain everything that is mentioned in the text and that pertains to an entity at the level of detail that it is decribed in the text. All information retains a link to where it comes from in the context variables, thus attributed properties will be tagged with *who* attributed them to this entity. Also, no attempt is made to resolve information beyond what is required to represent the text. For instance, beliefs will not be checked for their consistency and commonsense inferences will only be made if they are necessary to recover the coherence structure of the text.

The analysis here is restricted to capturing the utterances attributed to a discourse entity, which is stored in that person's profile[2]. The notion of "discourse entity" becomes unclear when abstract entities such as institutions are envolved. The Wall Street Journal corpus focuses on corporations, for instance, and we saw in Section 6.4.2 that the Wall Street Journal style elevates corporations and institutions to the rank of animate beings (or rather allows the appropriate metonymic extensions quite liberally). To identify that two utterances are attributed to the same source and therefore have to be stored in the same profile requires complex NP resolution. Some problems are:

1. Subsequent references may use a feature of the initial reference that is common knowledge.

2. Subsequent reference may introduce new information about the entity that is important to the immediate context.

3. A subsequent reference may introduce a new entity that speaks for the initial

[2]I will adopt here a way of referring to profiles that is not correct but simplifies the discussion. The impreciseness is that the primary information is not "stored" in the profile only. Rather, the primary information affixed with the context variables is represented in its propositional form at top level. A profile collects *pointers* to this information, just as the trace's temporally ordered discourse events are event variables that are linked to the events with pointers.

referent. This is not a proper case of "subsequent reference", but a related phenomenon.

4. Ambiguity in the reference may have to be resolved from the argumentative structure of the article.

The analysis of the text in the next chapter shows some of the complexities involved on an example. Here I will present the issues and representational solutions in isolation.

8.3.1 Accumulating Information

The purpose of profiles is to provide a place where information of very different kinds pertaining to a discourse entity can be stored. This collection can be made use of in different ways: when summarizing the article, information may be condensed; when extracting information, appropriate facets can be evaluated; when doing deep text analysis, this information can be used to start up commonsense inferencing.

Newspaper texts often introduce information about a discourse entity gradually, implicit in the lexicalization of the referring expressions of subsequent references. The incremental characterization usually adds up to a more comprehensive picture, where the underlying assumption is that all the characterizations are valid at the same time and at the time valid for the article. This holds for articles of the type sketched in the following text:

> **Emotional Agendas: Battle Over Abortion Heats Up in Wisconsin And a Few Other States**
>
> **EAU CLAIRE, Wis.** Barbara Lyons, the fast-talking executive director of Wisconsin Right to Life, is on the road, visiting radio and TV stations and talking to editors of local newspapers. It's her annual "media tour," and she's boosting one bill pending in the state legislature while roundly condemning another.
>
> . . .
>
> Mrs. Lyons and Mrs. Austin are the dairyland state's field marshals in the highly charged, deeply emotional battle over abortion.
>
> . . .
>
> By James M. Perry, Wall Street Journal 10/05/89

There are, however, articles where the assumption of cotemporaneity of the characteristics is clearly false, as in articles of the type illustrated in the following text:

> **Who's News:** Pacific Enterprises Chooses Ukropina As Chief Executive
> By Jeff Rowe
>
> Pacific Enterprises named its president, James R. Ukropina, as chairman and chief executive, succeeding Paul A. Miller and ending a century of family leadership at the utility holding company.
>
> Wall Street Journal, 10/5/89

It is of course the very meaning of the sentence to indicate a change in position for Ukropina, but other cases of characteristics changing over time may not be marked as clearly. It is therefore important to equip profiles with the possibility to *time stamp* certain properties (this does not entail having to commit to a specific form of temporal logic, because we do not attempt to infer a *complete* profile.) For the simplicity of the notation we do not require all information in a profile to carry a time stamp (also in line with MTR), but only provide a time stamp when it emerges from the text. The default assumption for characteristics without a time stamp is that they all hold at the time of the article.

8.3.2 Embedding Relations

One problem with the representation of information in profiles is the fact that different discourse entities may be connected in different ways. Individuals map very easily to profiles, one profile for each individual. Looking at abstract entities, such as *companies*, however, it becomes clear that profiles also have to stand in some structured relation to each other reflecting the relations of the entities, about which they contain information.

As demonstrated in Section 7.4, a frequent relationship made explicit in newspaper texts is that of *membership* or *representation*. There I showed that *employment* is a relationship that lets one person speak for a whole company.

To represent these *membership* or *part-of* relations between profiles, I will adopt the notation of nested boxes that has been used frequently for similar purposes [Fauconnier, 1985, Kamp, 1988, Ballim and Wilks, 1992]. The underlying intuition is that embedded

boxes are related to the embedding box in a relation similar to *is-a, part-of, member, representative-of.* Consider the employment relationship illustrated in Figure 8.3.

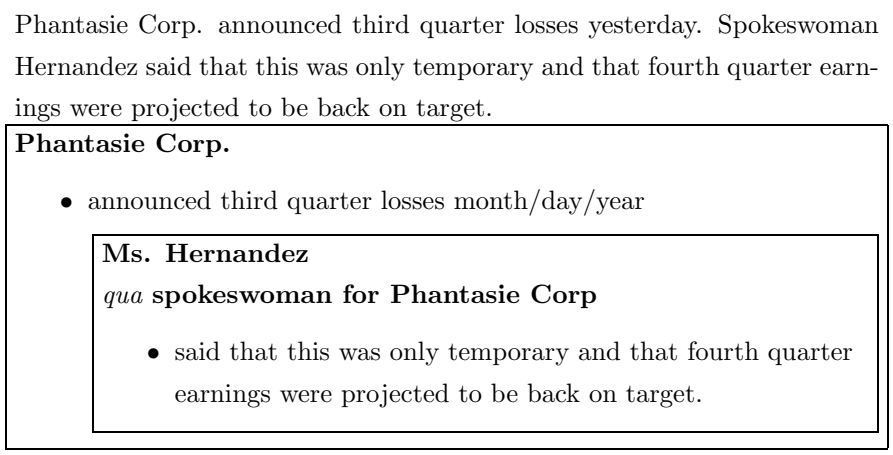


Figure 8.3: Employment relationship.

This simple text already shows representational difficulties:

(a) Hernandez is the spokeswoman for Phantasie Corp. In that capacity her statement holds for Phantasie Corp. and has to be embedded in that context.

(b) The announcement of third quarter losses was made by Phantasie Corp., thus is an instance of metonymy. We know (even on the level of lexical semantics) that there was an individual that made the announcement for Phantasie Corp. Should that fact be reflected at this level of the analysis (by introducing an unlabeled box)?

(c) Should the relationship between Phantasie Corp. and Hernandez (i.e. "spokeswoman") be represented twice, once as an access function from Phantasie Corp. to its spokesperson and once as an access function from Hernandez to her employer?

(a) reflects the problem that an individual plays many different roles: where would we put statements that Hernandez makes about her health club "Bigger Muscles", where she is treasurer? That information should clearly not be embedded in Phantasie Corp.'s profile!

It is important to remember that I am proposing a *text representation* scheme, not a knowledge representation scheme. Moreover, I am only considering newspaper articles

and in that context individuals usually appear only in one role at a time. I will therefore ignore this issue for now, making a mental note of it.

(b), in light of MTR, has in Figure 8.3 been answered tentatively with 'no'. Because a simple type of metonymy can be recovered from the lexical semantics, it can easily be recovered at this level, as well. To preserve the metonymy avoids having to deal with the problem of unifying boxes: it is most likely that Hernandez made the announcement as well, but the announcement could also have been made via press release by the PR office.

(c) addresses a problem that is somewhat related to (a), namely what the true representation of an 'embedding' should be: is there one object at "top level" for every profile and embeddings are implemented as pointer structures? Or should a profile be an object that is represented distributively, held together by pointers to the "mother profile"? I.e. is the box labeled *Hernandez* in Figure 8.3 a full (if not complete) profile or only one facet (the "spokeswoman–for–Phantasie–Corp." facet) of the profile? This question, like (a), will be noted but not answered here. For the purposes of characterizing the linguistic problems with reported speech, this issue can be ignored.

8.4 Complex Profile Structures

In order to devise a representation formalism for newspaper articles, it is important to consider what the salient features are that have to be captured. When representation formalisms are designed with an immediate goal in mind[3] domain knowledge can prime the knowledge representation effort. For a specific domain *templates* for several key (sublanguage) terms can be defined and serve as the goal output, reducing the analysis task to one of template filling.

The design of MTR explicitly avoids such a domain driven approach to designing the representation language. MTR rather attempts to substitute knowledge about language, language use, and the structure of newspaper articles for domain knowledge. Knowledge

[3]Two major research projects funded by DARPA take that route: extracting joint venture related information from the Wall Street Journal (TIPSTER) or extracting cases of terrorism from newspapers (MUC).

about the structure of newspaper articles of course constitutes domain knowledge of some sort, the difference is that it is applicable across the domains of joint ventures, micro chip design, and terrorist attacks.

8.4.1 'Aboutness' of the Article

Lundquist [Lundquist, 1989] argues that for polemic texts the initial sentence sets up a hypothesis (maybe indirectly) that the following article serves to argue for. In the Wall Street Journal, a similar effect of framing the argument can be observed for the first sentences. The first sentence of an article in the Wall Street Journal introduces in summarization the most salient information of the article, most importantly introducing the individuals, companies, institutions, or products that form the *topic* of the article. Compare:

(1) (a) Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC, was named a nonexecutive director of this British industrial conglomerate.

(b) A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago, researchers reported.

(c) Newsweek, trying to keep pace with rival Time magazine, announced new advertising rates for 1990 and said it will introduce a new incentive plan for advertisers.

(d) New England Electric System bowed out of the bidding for Public Service Co. of New Hampshire, saying that the risks were too high and the potential payoff too far in the future to justify a higher offer.

(e) The U.S., claiming some success in its trade diplomacy, removed South Korea, Taiwan and Saudi Arabia from a list of countries it is closely watching for allegedly failing to honor U.S. patents, copyrights and other intellectual-property rights.

(f) The Internal Revenue Service has threatened criminal sanctions against lawyers who fail to report detailed information about clients who pay them more than $10,000 in cash.

(g) The Transportation Department, responding to pressure from safety advocates, took further steps to impose on light trucks and vans the safety requirements used for automobiles.

These examples show that the first sentence already identifies a source for the information — usually this is the most salient or most important source; the most salient source is called the *root source*. As a heuristic we can assume that the root source is the

first source introduced into the text. The root source has a special significance: other sources are evaluated as confirming or contradicting the root source and the root source serves as an anchor for the profile building process.

The examples above also show that when there is a conflict of opinion or interest reported in the article, it is often mentioned in the first sentence already, setting up the *argumentative structure* for the article (see Section 8.4.3). Both parties might be mentioned explicitly, as is the case for (1c). In (1d) the opponent is not named but presupposed (*to bow out of the bidding* assumes that the bidding continues). (1e) only indirectly sets up a tension between the named countries, whereas in (1f) and (1g) the "opponent" is a particular profession.

This tendency to mention the major conflict in the article right in the first sentence is of great help in the determination of the *supporting group structure*.

8.4.2 Supporting Groups

When an article cites several people in favor of the same issue, possibly contrasting their statements with several other people's who are not in favor of this issue, there is often no employment or membership relation that would allow us to embed one profile into another. Yet for the sake of the argumentation in the article at hand, these profiles form a group, where the statement of one individual holds in some sense for the whole group. The statements *support* one another and I call the resulting grouping a *supporting group*.

To illustrate the notion, consider again the following paragraphs from the Wall Street Journal, October 5th, 1989:

> **Pacific Enterprises Chooses Ukropina As Chief Executive**
> By Jeff Rowe
>
> . . .
>
> Analysts said the naming of Mr. Ukropina represented a conservative move by an unusually conservative utility concern. Unlike some companies, Pacific Enterprises has "made no major errors moving outside their area of expertise," said Craig Schwerdt, an analyst with Seidler Amdec Securities Inc. in Los Angeles.
>
> "Each of the company's businesses are positioned to do well in the coming year," said Paul Milbauer, an analyst with C.J. Lawrence, Morgan Grenfell in New York.

> Most of the company's retail operations are in the fast-growing West, and the gas unit will benefit from tightening environmental regulations, he said. He added that more-stringent pollution controls are expected to increase demand for gas, which is relatively clean-burning.

The structure of this segment is roughly a claim (in the first sentence) followed by two supporting opinions, quoted directly from two independent sources. Note that there is no *employment* or *membership* relation between the two sources and the company. This supporting group is solely based on the gist of the opinions expressed. Supporting groups are represented in MTR as a simple box that groups a set of (supportive) boxes. Supporting groups may be labeled for reference.

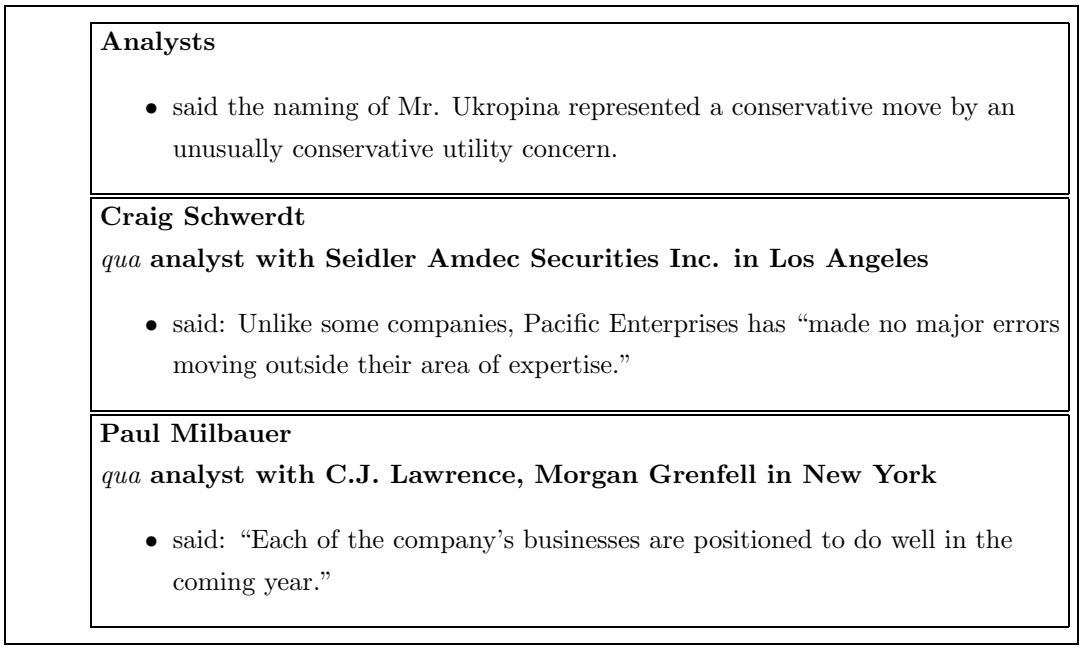


Figure 8.4: Supporting group of analysts.

Supporting groups are an important part of the text's structure. In those newspaper texts where different sources are mentioned it is important to be able to group the aruments being put forth. From the perspective of supporting group analysis there are three different forms of articles, namely

1. Articles that quote from a single source.

2. Articles that quote different sources to support a point.

3. Articles that report two or more different points of view, supporting each with quotes from different sources.

Supporting groups are tightly linked to the coherence structure. In the text fragment in Figure 8.4 it is the *elaboration* relationship between the sentences that establishes the supporting groups. The quoted material at the surface has nothing in common except being about the same company. But the structure (*claim* leads to *elaboration*), that is very frequent in newspaper articles, relates sentence three and four as being two independent instances elaborating the same claim. This is part of the coherence structure. Coherence structure in analysis does not emerge strictly prior to the structure of the supporting groups: If Paul Milbauer had predicted the company would "do poorly" rather than "do well", he would not have been in the same supporting group with Craig Schwerdt. This fact would have indicated that there cannot exist an elaboration relation between sentence three and the claim in sentence one, effectively determining coherence structure. For detail, see next section.

The next section illustrates the even tighter connectedness between supporting groups and coherence relations for articles of type 3, which I will call *argumentative articles*.

8.4.3 Opposing Supporting Groups

Chapter 4 demonstrated that newspaper articles often reflect implicitly the structure of a *stylized discourse* situation such as an interview, court proceedings, etc. It is important to consider the stylized discourse structure when determining supporting groups, because the stylized discourse provides stereotypical roles for the supporting groups. Court proceedings, for example, have three major roles, namely the plaintiff, the defendant, and the jury/judge. Each role has a corresponding supporting group and we expect both the defendant and the plaintiff supporting group to include material from at least two sources, namely an attorney and his or her client. Not only is the existence of these three supporting groups stipulated by the stylized discourse structure, but there is an underlying "coherence" relation between utterances made by members from the different supporting groups, namely *contradiction*. This is a valuable stipulation, because it can guide a commonsense reasoner to interpret the utterances until this

relation has been established.

The explicit supporting group structures are most helpful for structuring articles where opposing opinions are expressed by several sources. A variant of the Ukropina text will illustrate this point:

Variation I

(S_1) Pacific Enterprises named its president, James R. Ukropina, as chairman and chief executive, succeeding Paul A. Miller and ending a century of family leadership at the utility holding company.

(S_2') One analyst said the naming of Mr. Ukropina represented a conservative move by an unusually conservative utility concern. (S_3) Unlike some companies, Pacific Enterprises has "made no major errors moving outside their area of expertise," said Craig Schwerdt, an analyst with Seidler Amdec Securities Inc. in Los Angeles.

(S_4') "Each of the company's businesses are positioned to do poorly in the coming year," said Paul Milbauer, an analyst with C.J. Lawrence, Morgan Grenfell in New York. (S_5') Most of the company's retail operations are in the fast-growing West, and the gas unit will suffer from tightening environmental regulations, he said.

Variation I displays a different coherence structure from the original text. Both S_2' and S_4' are independent comments on S_1 elaborated by S_3 and S_5' respectively. The coherence structure for variation II is given in Figure 8.5. Note that in this example coherence structure cannot precede supporting group structure: the same information that allows us to make a distinction between *comment* and *elaboration* for S_4' also allows us to place the two statements into two opposing supporting groups.

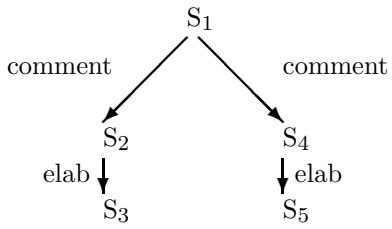


Figure 8.5: Coherence structure for variation I.

Variation I describes two opposing supporting groups. The only indication that

the two analysts belong to two opposing support groups lies in the content of their statements. This is a contrived example and in fact in real newspaper articles an opposition of this sort would not occur without any marker of opposition, such as *however* as an introduction, *another analyst* as explicit contrast to the first source, *insist* as reporting verb. Adding such a marker of opposition in S_4' immediately improves the flow of variation I.

Opposing supporting groups of analysts

"Good move"

One analyst

said the naming of Mr. Ukropina represented a conservative move by an unusually conservative utility concern.

Craig Schwerdt, analyst with Seidler Amdec Securities Inc. in Los Angeles

said: Unlike some companies, Pacific Enterprises has "made no major errors moving outside their area of expertise."

"Bad move"

Paul Milbauer, analyst with C.J. Lawrence, Morgan Grenfell in New York

said: "Each of the company's businesses are positioned to do poorly in the coming year."

said: Most of the company's retail operations are in the fast-growing West, and the gas unit will suffer from tightening environmental regulations.

Coherence relations are a major tool in establishing text structure (cf.[Hobbs, 1982, Polanyi and Scha, 1984, Polanyi, 1987, Scha and Polanyi, 1988, Cohen, 1987, Mann and Thompson, 1986].) Argumentative articles can introduce two coherence relations for every sentence, namely a statement can serve as both an elaboration of a previously made statement of the same support group and as a refutation or contradiction of a statement made previously by a member of the opposing support group. Thus the coherence structure is not linear (i.e. a path through an utterance tree) as suggested

in the literature mentioned above, but at least two dimensional (i.e. the coherence structure produced is a tangled hierarchy or a general graph). This, I claim, is in fact the distinguishing feature between argumentative and expository texts; in the latter care is taken to *produce* a linear coherence structure for clarity, where in the former the complexity of the argument is reflected in a tangled text structure.

To elaborate on the notion of argumentative text and its coherence relations is beyond the scope of this dissertation. But it is important to acknowledge that in argumentative articles the coherence relation of a sentence to its preceeding context can be construed in more than one way, allowing members of different supporting groups to *reinterpret* them. I.e. the attorney of the defendant can reinterpret the statement of a witness for the plaintiff as not harmful to his or her client, where the attorney for the plaintiff will interpret that same statement as very damaging for the case of the defendant. The intricacies of the court proceedings are rarely reported in detail in newspaper articles, where summaries of the outcome are more relevant.

Another scenario that proves difficult for the notion of supporting groups is the case of a round table discussion with different experts. Here, different supporting groups might be formed on every topic discussed. This raises again the issue of distributed representation of profiles. Ultimately, it seems necessary to represent profiles in a distributed manner. Yet the comment made in Section 8.3.2 is valid here, too: for a majority of newspaper articles this problem does not arise, and I will therefore treat the profiles here as single boxes for simplicity.

8.4.4 Tripartite Representation

Trace and profiles are only two aspects of text representation, in fact two aspects that could theoretically be derived from a representation of the propositional contents of the sentences connected by the cohesion and coherence relations that are implicit in the text sequence or even explicitly marked. I have argued here that it is profitable to extract such specialized representational aspects and to represent them separately. Yet they cannot replace the basic propsitional structure.

One way to preserve the basic information and the flow of the text is to keep a record of every sentence in sequence. This works well for simple sentences. The advantage of a

tripartite representation (Trace, Profiles, and Coherence Structure) is the possibility to summarize a text efficiently in three very different ways: From the point of view of one particular discourse entity (profile structure), chronologically (trace), and along the lines of the original text (coherence structure). For information retrieval the importance of the three separate structures is one of efficiency: A proposition in the text that matches a query can easily be cross-indexed and its temporal context as well as the point of view can be established without having to analyze the whole text.

8.4.5 Summary

MTR is a text representation scheme that combines three devices, coherence structure, trace, and profile structure, to record the structure of a text faithfully yet in a concise way. Because the three devices represent different aspects of the text their individual structures can be kept simple and can guide different uses of the text analysis.

Profile structure, a novel concept, is capable not only to connect divers aspects attributed to a discourse entity, but also to reflect the argumentative structure of a text. Providing supporting and opposing supporting group structures allows to interpret isolated utterances in the correct global context.

Chapter 9

Text Analysis

This chapter demonstrates how the interpretation process proceeds and what knowledge structures are involved in minimal text representation. The text chosen to demonstrate the mechanism is both longer and more complex than is usual for demonstration purposes. I chose this level of complexity, however, to show in an integrated example how the representation mechanism and the default heuristics work for newspaper articles.

The previous chapters introduced reported speech as a pervasive phenomenon in newspaper articles that requires a specialized interpretation mechanism which could in part be triggered by the lexical semantics of the reporting verb, in part by structural considerations. With the additional help of a representation scheme that addresses the special requirements for complex reported speech contexts, MTR, I will here show that common belief sets expressed in newspaper articles can largely be recovered by determining the article's thematic structure of *opposition* and *support*. The thesis of this chapter is:

1. The beliefs of individuals can be captured by how their behavior is represented in the text.

2. This behavior is expressed in the reporting clause and in the lexicalization of the source and the reporting verb.

3. The individuals which are portrayed as sources are logically related as either sup-

porting or opposing each others view. Supporting relationships can often be inferred from semantically linked lexicalizations.

4. This thematic structure can be captured making use of MTR structures as relating to lexical semantic representation in a GL semantics.

The text I will analyze in depth is the following:

> **Regulators Cite Delays and Phone Bugs**
>
> Regional thrift regulators charged that phone calls were bugged during the examination of Lincoln Savings & Loan Association and that federal officials in Washington delayed Lincoln's seizure until its $2.5 billion cost made it the most expensive thrift failure ever.
>
> "Clearly, we were shot in the back . . . as we battled to protect the taxpayers," said William Black, acting district counsel for the San Francisco region of thrift regulators.
>
> In a day of extraordinary testimony before the House Banking Committee, officials from the San Francisco Office of Thrift Supervision also testified that a Big Eight accounting firm, Arthur Andersen & Co., participated in back-dating loan documents and that the Washington officials even agreed in one document not to prosecute Lincoln over certain infractions.
>
> . . .
>
> Witnesses said the room in which California examiners were auditing Lincoln this past spring was bugged. According to a government memo, regulators who took control of the thrift in August discovered that a phone line in Lincoln's headquarters in Irvine, Calif., had been "compromised" to allow calls to be monitored on other extensions. It didn't say who was monitoring the calls, but said the matter was turned over to the Federal Bureau of Investigation.
>
> A Lincoln spokesman said its management "never authorized or participated in any bugging of anyone."
>
> . . .
>
> The regional officials said Washington's chief thrift regulator, Danny Wall, and his principal lieutenants repeatedly ignored warnings that Lincoln was being operated in a reckless manner, certain to cause its failure. They also accused Mr. Wall of holding improper meetings with Lincoln officials, while refusing to listen to field examiners.
>
> Members of the Banking Committee generally received the regional officials as bureaucratic heroes and sharply criticized their Washington bosses, who had relieved them of their responsibility for Lincoln two years before the thrift's parent filed for bankruptcy-law protection.
>
> Committee Chairman Henry Gonzalez (D., Texas) said Mr. Wall "willingly cut the legs out from under his regulatory troops in the middle of the battle," and renewed

his call for Mr. Wall to step aside. Rep. Joseph Kennedy (D., Mass.) said, "The higher up in the regulatory process, the more corrupt it would appear."

Mr. Wall is scheduled to testify Nov. 7, and his agency said that all sides of the issue should be heard before drawing conclusions. Separately, other Washington thrift officials disputed the regional examiners' statement that they had called specifically for the seizure of Lincoln in 1987, saying that it was only one option they presented.

The regional officials also said that Arthur Andersen backdated data to support loans that were made with no underwriting standards. Over two years, until April 1986, $1 billion in loans were approved, even though Lincoln had no written loan standards, said Mike Patriarca, acting principal supervisory agent with the Office of Thrift Supervision. Fifty-two loans were made in March 1986, he said, and none had credit reports or other background work completed.

"At a later time, the files looked good because they had been stuffed," Mr. Patriarca said. Leonard Bickwit, a Washington attorney for Lincoln, conceded that some memos had been written after the fact. He said that "memorialization of underwriting activities that had been undertaken at an earlier time" did occur, but that Lincoln believed it adhered to lending standards superior to the industry average.

A spokesman for Arthur Andersen denied any improprieties, adding, "At the request of our then-client, we provided staff personnel to work for a limited period of time under the direction of client personnel to assist them in organizing certain files."

...

By Paulette Thomas and Brooks Jackson, Wall Street Journal, 10/27/89

9.1 Separating Information

The first step in representing a complex text according to the principles of MTR is to separate the three sentence components for reported speech sentences: the source, the reporting verb and modifications, and the complement.[1] Given a robust (partial) parser (cf. [McDonald, 1991]) this step can be done automatically.

The separation of information presented in tabular form gives the following picture (note that a horizontal line indicates a sentence boundary, two horizontal lines a paragraph boundary):

[1]Sentences that are not reported speech will stay unanalyzed here.

Regional thrift regulators	charged that	phone calls were bugged during the examination of Lincoln Savings & Loan Association
	and that	federal officials in Washington delayed Lincoln's seizure until its $2.5 billion cost made it the most expensive thrift failure ever.
William Black, acting district counsel for the San Francisco region of thrift regulators	said	"Clearly, we were shot in the back . . . as we battled to protect the taxpayers."

In a day of extraordinary testimony before the House Banking Committee,

officials from the San Francisco Office of Thrift Supervision	also testified that	a Big Eight accounting firm, Arthur Andersen & Co., participated in back-dating loan documents
	and that	the Washington officials even agreed in one document not to prosecute Lincoln over certain infractions.
Witnesses	said	the room in which California examiners were auditing Lincoln this past spring was bugged.
a government memo	According to	regulators who took control of the thrift in August discovered that a phone line in Lincoln's headquarters in Irvine, Calif., had been "compromised" to allow calls to be monitored on other extensions.
It	didn't say but said	who was monitoring the calls, the matter was turned over to the Federal Bureau of Investigation.
A Lincoln spokesman	said	its management "never authorized or participated in any bugging of anyone."

. . .

The regional officials	said	Washington's chief thrift regulator, Danny Wall, and his principal lieutenants repeatedly ignored warnings that Lincoln was being operated in a reckless manner, certain to cause its failure.

They	also accused	Mr. Wall of holding improper meetings with Lincoln officials, while refusing to listen to field examiners.
Members of the Banking Committee	generally received the regional officials as bureaucratic heroes	
	and sharply criticized	their Washington bosses, who had relieved them of their responsibility for Lincoln two years before the thrift's parent filed for bankruptcy-law protection.
Committee Chairman Henry Gonzalez (D., Texas)	said	Mr. Wall "willingly cut the legs out from under his regulatory troops in the middle of the battle,"
	and renewed his call	for Mr. Wall to step aside.
Rep. Joseph Kennedy (D., Mass.)	said,	"The higher up in the regulatory process, the more corrupt it would appear."
	Mr. Wall is scheduled to testify Nov. 7,	
his agency	said	that all sides of the issue should be heard before drawing conclusions.
Separately, other Washington thrift officials	disputed	the regional examiners' statement that they had called specifically for the seizure of Lincoln in 1987,
	saying that	it was only one option they presented.
The regional officials	also said that	Arthur Andersen backdated data to support loans that were made with no underwriting standards.
Mike Patriarca, acting principal supervisory agent with the Office of Thrift Supervision	said	over two years, until April 1986, $1 billion in loans were approved, even though Lincoln had no written loan standards.
He	said	fifty-two loans were made in March 1986
	and	none had credit reports or other background work completed.
Mr. Patriarca	said	"At a later time, the files looked good because they had been stuffed"

Leonard Bickwit, a Washington attorney for Lincoln,	conceded that	some memos had been written after the fact.
He	said that	"memorialization of underwriting activities that had been undertaken at an earlier time" did occur,
	but that	Lincoln believed it adhered to lending standards superior to the industry average.
A spokesman for Arthur Andersen	denied	any improprieties
	adding	"At the request of our then-client, we provided staff personnel to work for a limited period of time under the direction of client personnel to assist them in organizing certain files."

It is interesting to note that most sentences in this article follow exactly the schema <source><reporting verb><complement> and can be read left to right in this table.

This table was constructed based on the syntactic categories noun phrase, verb, and complement and it suggests a different control structure for the following text analysis. Leaving the complement unanalyzed for the moment, the utterances can be placed into profiles based on the content of the source noun phrase and the coherence of the reporting verb. This partial analysis can then in turn facilitate the analysis of the complements within the local context of the profiles. I will assume this strategy here for the following discussion.

9.2 Profiling

The next step in builing a representation of the text is the process of collecting the information in profiles, called *profiling* for short. The major task of profiling is reference resolution and the interpretation of complex NPs. Profiling is harder, if there are many entities involved.

PROFILING(*source*, *reporting-verb*, *complement*, *list-of-profiles*)

if $\exists\ p \in$ *list-of-profiles* such that
refers-to(name(p)), x & refers-to(*source*, x)

then add-to-profile(p, *reporting-verb*, *complement*)
else make-profile(*source*, *reporting-verb*, *complement*)

The profiles for the text under consideration are:

Regional thrift regulators from the San Francisco Office of Thrift Supervision

- charged that phone calls were bugged during the examination of Lincoln Savings & Loan Association
- charged that federal officials in Washington delayed Lincoln's seizure until its $2.5 billion cost made it the most expensive thrift failure ever.

William Black, acting district counsel for the San Francisco region of thrift regulators

- said "Clearly, we were shot in the back . . . as we battled to protect the taxpayers."

- testified that a Big Eight accounting firm, Arthur Andersen & Co., participated in back-dating loan documents
- testified that the Washington officials even agreed in one document not to prosecute Lincoln over certain infractions.
- said Washington's chief thrift regulator, Danny Wall, and his principal lieutenants repeatedly ignored warnings that Lincoln was being operated in a reckless manner, certain to cause its failure.
- accused Mr. Wall of holding improper meetings with Lincoln officials, while refusing to listen to field examiners.
- said that Andersen backdated data to support loans that were made with no underwriting standards.

Mike Patriarca, acting principal supervisory agent with the Office of Thrift Supervision

- said over two years, until April 1986, $1 billion in loans were approved, even though Lincoln had no written loan standards.
- said fifty-two loans were made in March 1986, and none had credit reports or other background work completed.
- said "At a later time, the files looked good because they had been stuffed"

Witnesses

- said the room in which California examiners were auditing Lincoln this past spring was bugged.

A government memo

- 'according to which' regulators who took control of the thrift in August discovered that a phone line in Lincoln's headquarters in Irvine, Calif., had been "compromised" to allow calls to be monitored on other extensions.
- didn't say who was monitoring the calls,
- said the matter was turned over to the Federal Bureau of Investigation.

A Lincoln spokesman

- said its management "never authorized or participated in any bugging of anyone."

Members of the Banking Committee

- sharply criticized their [the regional thrift regulator's –sb] Washington bosses, who had relieved them [the regional thrift regulators –sb] of their responsibility for Lincoln two years before the thrift's parent filed for bankruptcy-law protection.

Committee Chairman Henry Gonzalez (D., Texas)

- said Mr. Wall "willingly cut the legs out from under his regulatory troops in the middle of the battle,"
- renewed his call for Mr. Wall to step aside.

Rep. Joseph Kennedy (D., Mass.)

- said "The higher up in the regulatory process, the more corrupt it would appear."

Mr. Wall

- is scheduled to testify Nov. 7

Mr. Wall's agency

- said that all sides of the issue should be heard before drawing conclusions.

Other Washington thrift officials

- disputed the regional examiners' statement that they had called specifically for the seizure of Lincoln in 1987
- saying that it was only one option they presented.

Leonard Bickwit, a Washington attorney for Lincoln

- conceded that some memos had been written after the fact.
- said that "memorialization of underwriting activities that had been undertaken at an earlier time" did occur
- said that Lincoln believed it adhered to lending standards superior to the industry average.

A spokesman for Arthur Andersen

- denied any improprieties
- adding "At the request of our then-client, we provided staff personnel to work for a limited period of time under the direction of client personnel to assist them in organizing certain files."

There are some interesting decisions reflected in this profile structure. For instance, the NPs "a spokesman for XYZ" have resulted in a profile for "a spokesman of XYZ", not a profile for *XYZ* and embedded a profile for the spokesman. This indicates the conservative structure of the interpretation process. Also, some discourse markers, such as *and* and *but* have been omitted, obfuscating the coherence relations between clauses. And finally it is obvious that the profile structure cannot preserve the argumentative flow of the underlying article. These coherence structure links (and also temporal re-

lationships lost in the profiles) are of course represented (and cross indexed) in the Coherence Structure and the Trace respectively.

9.3 Grouping

Next, the profiles have to be *grouped* into supporting groups. This process requires the analysis of the complements and the reporting context in order to determine supporting group structure in addition to the reference resolution and analysis of comlpex NPs assumed for the profiling.

The grouping process is very complex and requires in the worst case full text analysis in order to be successful. Applying some heuristics gained from the discussion in earlier chapters can in certain cases cut down the complexity considerably.

The first heuristic to be applied is a similarity judgement on source descriptions. A profile can be placed safely into a supporting group if the source description of the profile shows some strong coherence link with another profile (preferrably the root source) in the supporting group.

This heuristic holds in the Thrift Regulation article for example for the profiles called "A Lincoln spokesman" and "Leonard Bickwit, a Washington attorney for Lincoln".

The similarity of the names of the profiles can of course exist on many different levels and is as difficult as placing an individual utterance within a profile.

If the name of the profile does not suggest a grouping with exiting supporting groups, the reporting verb is the next choice. Reporting verbs may encode positive or negative coherence with the previous sentence and indeed we find in the article that *charge, testify* and *accuse* are verbs used for the plaintiff supporting group and *dispute, concede*, and *deny* for the defendant supporting group (more on this discussion in Section 9.4). Especially in conjunction with a stylized discourse script on Congressional hearings this coherence information places the appropriate supporting groups. This is of course not as simple as it sounds: a plaintiff could well deny claims by the defendant, thus upsetting the simpleminded application of this heuristic. The whole profile has to support the choice of supporting group.

The third step is to consider general coherence markers, especially markers of opposition such as *but* or *however*. Again, these do not in isolation determine which supporting group a profile belongs to. In articles with only two opposing supporting groups they provide very strong evidence. In other cases they rather serve to eliminate one supporting group from the set of choices.

Lastly, the content of the complement clause has to be considered in order to find out which argument position it supports. The four steps have to be seen as a heuristic ordering of steps that cannot totally determine the grouping, but that can suggest a grouping.

The algorithm cannot be given in detail here — the task of finding coherence links between source descriptions or reporting verbs is much too complicated in the general case. As the discussion above points out, exceptions to the general and complex case are not infrequent and it is the simpler case that this ordering of heuristics tries to take advantage of.

The outline of the algorithm is as follows:

GROUPING(*profile*, *list-of-supporting-groups*)

if ∃ a supporting group $s \in$ *list-of-supporting-groups* &
∃ a profile $p \in s$ such that
name(*profile*) is-similar-to name(p)
then add(*profile*, s)
else **if** ∃ a supporting group $s \in$ *list-of-supporting-groups* &
∃ a profile $p \in s$ such that
reporting-verbs(*profile*) coher-with reporting-verbs(s)
then add(*profile*, s)
else **if** ∃ a supporting group $s \in$ *list-of-supporting-groups* &
∃ a coherence relation c between a sentence $S_1 \in$ *profile* and
some sentence S_2 in s such that
marks-opposition(c)
then incompatible(*profile*, s)
else **if** ∃ a supporting group $s \in$ *list-of-supporting-groups* &
∃ a profile $p \in s$ such that
argumentation-in(*profile*) support argumentation-in(s)
then add(*profile*, s)

Note that a grouping into supporting groups does not have to be complete. Any supporting group structure can be of great help for the further exploitation of the representation. The philosophy behind MTR, however, prevents that a minimal number of supporting groups has to be formed.

Let me present the supporting group structure for the article and then discuss some of the particular coherence relations found in this text afterwards.

Plaintiff

Regional thrift regulators from the San Francisco Office of Thrift Supervision

- charged that phone calls were bugged during the examination of Lincoln Savings & Loan Association
- charged that federal officials in Washington delayed Lincoln's seizure until its $2.5 billion cost made it the most expensive thrift failure ever.

William Black, acting district counsel for the San Francisco region of thrift regulators

- said "Clearly, we were shot in the back . . . as we battled to protect the taxpayers."

- testified that a Big Eight accounting firm, Arthur Andersen & Co., participated in back-dating loan documents
- testified that the Washington officials even agreed in one document not to prosecute Lincoln over certain infractions.
- said Washington's chief thrift regulator, Danny Wall, and his principal lieutenants repeatedly ignored warnings that Lincoln was being operated in a reckless manner, certain to cause its failure.
- accused Mr. Wall of holding improper meetings with Lincoln officials, while refusing to listen to field examiners.
- said that Andersen backdated data to support loans that were made with no underwriting standards.

Plaintiff — continued

Regional thrift regulators — continued

Mike Patriarca, acting principal supervisory agent with the Office of Thrift Supervision

- said over two years, until April 1986, $1 billion in loans were approved, even though Lincoln had no written loan standards.
- said fifty-two loans were made in March 1986, and none had credit reports or other background work completed.
- said "At a later time, the files looked good because they had been stuffed"

Witnesses

- said the room in which California examiners were auditing Lincoln this past spring was bugged.

A government memo

- 'according to which' regulators who took control of the thrift in August discovered that a phone line in Lincoln's headquarters in Irvine, Calif., had been "compromised" to allow calls to be monitored on other extensions.
- didn't say who was monitoring the calls,
- said the matter was turned over to the Federal Bureau of Investigation.

Defendant

A Lincoln spokesman

- said its management "never authorized or participated in any bugging of anyone."

Defendant — continued

Mr. Wall

- is scheduled to testify Nov. 7

Mr. Wall's agency

- said that all sides of the issue should be heard before drawing conclusions.

Other Washington thrift officials

- disputed the regional examiners' statement that they had called specifically for the seizure of Lincoln in 1987
- saying that it was only one option they presented.

Leonard Bickwit, a Washington attorney for Lincoln

- conceded that some memos had been written after the fact.
- said that "memorialization of underwriting activities that had been undertaken at an earlier time" did occur
- said that Lincoln believed it adhered to lending standards superior to the industry average.

A spokesman for Arthur Andersen

- denied any improprieties
- adding "At the request of our then-client, we provided staff personnel to work for a limited period of time under the direction of client personnel to assist them in organizing certain files."

Committee

Members of the Banking Committee

- sharply criticized their [the regional thrift regulator's –sb] Washington bosses, who had relieved them [the regional thrift regulators –sb] of their responsibility for Lincoln two years before the thrift's parent filed for bankruptcy-law protection.

Committee Chairman Henry Gonzalez (D., Texas)

- said Mr. Wall "willingly cut the legs out from under his regulatory troops in the middle of the battle,"
- renewed his call for Mr. Wall to step aside.

Rep. Joseph Kennedy (D., Mass.)

- said "The higher up in the regulatory process, the more corrupt it would appear."

Based on a rough "Congressional Committee Hearing" script, the division of the profiles here provides a frame for the argumentative structure of the article. This argumentative structure is important in several respects. First, it makes clear that two branches of the same institution are in fact in different supporting groups. This fact could easily be lost if a simplistic reasoner were to judge from the institutional structure that two branches of the same institution did not differ in their opinions.

Secondly, it supplies the proper set of presupposed and implied information: The utterances reported for members of the supporting group "Defendant" can only be meaningfully interpreted in the context of the allegations of members of the supporting group "Plaintiff". The underlying structure of the text is that of a (court) hearing, here a hearing of the banking committee.

It is however obvious, that the supporting group structure can only supply partial evidence for the argumentative structure, since the very grouping into supporting groups results in a loss of the coherence or text structure of the original text. This structure, as mentioned in Chapter 8, has to be represented independently. The interaction of both structures will allow for efficient manoevering through the text representation. I will

not sketch text structure here, see [Polanyi, 1987] for one approach.

9.4 Lexical Semantics and Coherence

This section illustrates what the lexical semantics of different words contributes to the construction of profiles and supporting groups in this article. First, let us consider the subject NPs, listed in the leftmost column of the separated text in Section 9.1[2].

9.4.1 Source NPs

The most interesting and varied supporting group is "Plaintiff". The general profile for "regional thrift regulaters" contains many utterances. The source NPs can all be dereferenced based on one single "key", namely *regional*. *Regional* here contrasts with *Washington* and *federal* and is used as a keyword to distinguish the two bureaus of the same institution, that in fact fall into different supporting groups. This is an important observation that might go unnoticed in a keyword search strategy for information retrieval, which might focus on *thrift regulators* as the key. *Regional* is not sufficient as a keyword — in the third sentence the discourse entity is referred to as *officials from the San Francisco Office of Thrift Supervision*, which can be dereferenced because in the previous sentence the entity *William Black* was introduced as *acting district counsel for the San Francisco region of thrift officials*. Then the contrast to *federal* and *Washington officials* is obvious.

In order to properly make this distinction, a text understanding system has to have a rudimentary model of American political geography, namely that *Washington* city is the capital and seat of most *federal* agencies, that then have *regional* offices throughout the country. Even a linguisticall oriented system like MTR has to incorporate such world knowledge because it is implicit in the lexical semantics of the words *as they are*

[2]In the previous sections in this chapter the complement clauses stood unanalyzed. This choice illustrates that such a partial analysis already provides important insight into the structure of the text. In a full analysis, of course, all text is analyzed and the profiles are composed incrementally from information found in NPs both of the matrix and the complement. The shortcomings of the "partial" analysis can be seen in the "Defendant" supporting group: Mr. Wall's profile there does not contain any information about his person, which was introduced earlier in the text in a complement clause.

used. A system that does not have this knowledge might adequately dereference the descriptive NPs based on recency considerations or topic–focus structure, but it will not be able to truly "understand" the oppositions in this particular text.

Notice also that *Washington* here is used coercively, not metonymically as discussed in Section 6.4.3, but with similar results. Introduced as *federal officials in Washington* in complement position in the first sentence, they are referred to subsequently as *Washington officials*, *Washington bosses*, and *other Washington thrift officials*. *Washington officials* is a *contextual* coercion, where some contextually given information about the entity fills the semantic argument position for *official*. In Section 5.4 we saw that *official* requires the specification of an institution or organization for which the *official* works. I called this position a semantic argument, that does not have to be realized syntactically as an argument but has to be inferentially derivable. Here the semantic argument ought to be filled literally with something like *the federal Office of Thrift Supervision in Washington* resulting in the monster NP *the Washington based federal Office of Thrift Supervision official* which is hard to comprehend, although well-formed. Language economy and in fact semantic clarity require that the modifier be abbreviated with the most distinguishing term, here *federal* or *Washington* in contrast to *regional* and *San Francisco*.

9.4.2 Reporting Verbs

Another interesting lexical choice in this article concerns the reporting verbs used in different supporting groups. In the "Plaintiff" group we find *charge, testify, accuse*, contrasting with *dispute, concede, deny* in the "Defendant" group. The lexical semantics of these reporting verbs confirms the coherence of the supporting groups. The first difference to discuss is the semantic dimension introduced in Section 3.3.1 (see Figure 3.5): *charge, testify,* and *accuse* imply *new* material, whereas *dispute, concede,* and *deny* implicitly refer to a proposition, moreover to an *accusation*, which is paraphrased in the complement.

On another level, *dispute, concede,* and *deny* are general terms that are not bound to the stylized discourse script for (court) hearings. But *accuse* and *charge* imply three parties involved, the perpetrator, the accuser, and an audience (which is implicitly

of some authority in this matter). Note that the audience can be identical to the perpetrator, making his conscience the authority appealed to. *Testify*, on the other hand, envolves two parties (the witness and the audience) and an event; it, too, evokes a (court) hearing situation.

The previous sections have shown that lexical semantics provides coherence relations already on a very superficial level of analysis. These coherence relations strengthen the usefulness of constructing a supporting group structure for articles that report two or more points of view. Lexical coherence can also be exploited for the grouping process.

9.5 Summary

This chapter illustrated on a complex example that profiles and supporting groups are a useful tool in the assessment of the argumentative structure of an article. I have discussed heuristics that can facilitate the process of building these structures but much work remains to be done before it can be automated. Allowing the supporting groups to be build incrementally and allowing partial representation we can, however, exploit supporting group structure where it can be obtained at reasonable cost.

Bibliography

[Amsler, 1980] R.A. Amsler. *The Structure of the Merriam-Webster Pocket Dictionary*. PhD thesis, University of Texas, 1980.

[Anderson, 1986] L.B. Anderson. Evidentials, paths of change, and mental maps: Typologically regular asymmetries. In *Evidentiality: The Linguistic Coding of Epistemology*. Ablex, Norwood, 1986.

[Anick and Bergler, 1991] P. Anick and S. Bergler. Lexical structures for linguistic inference. In *Proceedings of the First SIGLEX Workshop on Lexical Semantics and Knowledge Representation*, Berkeley, June 1991.

[Anick and Pustejovsky, 1990] P. Anick and J. Pustejovsky. An application of lexical semantics to knowledge acquisition from corpora. In *Proceedings of the 13th International Conference on Computational Linguistics, Helsinki, 1990*, 1990.

[Bach and Harnish, 1979] K. Bach and R.M. Harnish. *Linguistic Communication and Speech Acts*. MIT Press, Cambridge, U.S., 1979.

[Ballim and Wilks, 1992] A. Ballim and Y. Wilks. *Artificial Believers*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1992.

[Ballim *et al.*, 1991] A. Ballim, Y. Wilks, and J.A. Barnden. Belief ascription, metaphor, and intensional identification. *Cognitive Science*, 15:133–171, 1991.

[Banfield, 1982] A. Banfield. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge & Kegan Paul, Boston, 1982.

[Bergler and Pustejovsky, 1990] S. Bergler and J. Pustejovsky. Temporal reasoning from lexical semantics. In B. Endress-Niggemeyer, T. Herrmann, A. Kobsa, and D. Rösner, editors, *Interaktion und Kommunikation mit dem Computer*, pages 145–154. Springer Verlag, 1990.

[Bergler, 1991] S. Bergler. The semantics of collocational patterns for reporting verbs. In *Proceedings of the Fifth European Conference of the Association for Computational Linguistics*, pages 216–221, Berlin, Germany, April 1991.

[Bernth, 1990] A. Bernth. Treatment of anaphoric problems in referentially opaque contexts. In R. Studer, editor, *Natural Language and Logic*, Lecture Notes in Artificial Intelligence, pages 1–25. Springer Verlag, Berlin, 1990.

[Berube, 1987] M.S. Berube, editor. *The American Heritage Dictionary*. Dell Publishing Co., Inc., New York, 1987. Paperback edition based on the Second College Edition, 1983, Houghton Mifflin Co.

[Boguraev and Briscoe, 1989] B. Boguraev and T. Briscoe, editors. *Computational Lexicography for Natural Language Processing*. Longman, Harlow, Essex, 1989.

[Bruce, 1978] B. Bruce. What makes a good story? Reading Education Report 5, Bolt, Beranek and Newman, Cambridge, Massachusetts, 1978.

[Chafe, 1986] W. Chafe. Evidentiality in english conversation and academic writing. In W. Chafe and J. Nichols, editors, *Evidentiality: The Linguistic Coding of Epistemology*. Ablex, New Jersey, 1986.

[Church and Hanks, 1990] K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 1990.

[Church, 1988] K.W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas*, 1988.

[Clark and Gerrig, 1990] H. Clark and R. Gerrig. Quotations as demonstrations. *Language*, 66(4):764–805, 1990.

[Cohen, 1987] R. Cohen. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13(1–2), 1987.

[Davidson, 1968] D. Davidson. On saying that. *Synthèse*, 19:130–146, 1968. Reprinted in D. Davidson, *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford, 1984.

[Davidson, 1979] D. Davidson. Quotation. *Theory and Decision*, 11:27–40, 1979. Reprinted in D. Davidson, *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford, 1984.

[Dowty *et al.*, 1981] D.R. Dowty, R.E. Wall, and S. Peters. *Introduction to Montague Semantics*. Reidel Publishing Co., Dordrecht, Holland, 1981.

[Fano, 1961] R. Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, Mass, 1961.

[Fass, 1988] D. Fass. An account of coherence, semantic relations, metonymy, and lexical ambiguity resolution. In S. Small, G. Cottrell, and M. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. Morgan Kaufmann Publishers, San Mateo, 1988.

[Fauconnier, 1985] G. Fauconnier. *Mental Spaces*. Bradford Books. MIT Press, Cambridge, MA, 1985.

[Fillmore, 1968] C. Fillmore. The case for case. In E. Bach and R.T. Harms, editors, *Universals in Linguistic Theory*. Holt and Rinehart, New York, 1968.

[Frajzyngier, 1985] Z. Frajzyngier. Truth and the indicative sentence. *Studies in Language*, 9(2):243–254, 1985.

[Francis and Kučera, 1982] W. Francis and H. Kučera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin Company, Boston, MA, 1982.

[Gazdar, 1979] G. Gazdar. *Pragmatics: Implicature, Presupposition, and Logical Form*. Academic Press, New York, 1979.

[Gazdar, 1981] G. Gazdar. Speech act assignment. In A. Joshi, B. Webber, and I. Sag, editors, *Elements of Discourse Understanding*. Cambridge University Press, Cambridge, U.K., 1981.

[Geach, 1972] P.T. Geach. Quotation and quantification. In *Logic Matters*. Blackwell, Oxford, 1972.

[Givón, 1982] T. Givón. Evidentiality and epistemic space. *Studies in Language*, 6(1), 1982.

[Grice, 1957] H.P. Grice. Meaning. *Philosophical Review*, 66, 1957.

[Grice, 1967] H.P. Grice. Logic and conversation. Unpublished MS. of the William James Lectures, Harvard University, 1967.

[Grimshaw, 1990] J. Grimshaw. *Argument Structure*. MIT Press, Cambridge, USA, 1990.

[Grishman *et al.*, 1986] R. Grishman, L. Hirschman, and Ngo Thanh Nhan. Procedures for sublanguage selectional patterns: Initial experiments. *Computational Linguistics*, 12(3), 1986.

[Grosz and Sidner, 1986] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 1986.

[Hand, 1991] M. Hand. On saying that again. *Linguistics and Philosophy*, 14:349–365, 1991.

[Hindle, 1983] D. Hindle. User manual for Fidditch. Memorandum 7590-142, Naval Research Laboratory, 1983.

[Hindle, 1990] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th Meeting of the Association for Computational Linguistics*, 1990.

[Hinrichs, 1986] E. Hinrichs. Temporal anaphora in discourses of English. *Linguistics and Philosophy*, 9, 1986.

[Hintikka, 1969] J. Hintikka. Semantics for propositional attitudes. In J.W. et al. Davis, editor, *Philosophical Logic*. Reidel, Dordrecht, 1969.

[Hobbs *et al.*, 1987] J.R. Hobbs, W. Croft, T. Davies, D. Edwards, and K. Laws. Commonsense metaphysics and lexical semantics. *Computational Linguistics*, 13(3–4), 1987.

[Hobbs, 1979] J.R. Hobbs. Coherence and coreference. *Cognitive Science*, 3(1):67–90, 1979.

[Hobbs, 1982] J.R. Hobbs. Towards an understanding of coherence in discourse. In W. Lehnert and M. Ringle, editors, *Strategies for Natural Language Processing*. Lawrence Erlbaum Associates, New Jersey, 1982.

[Hobbs, 1985] J.R. Hobbs. Ontological promiscuity. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, 1985.

[Hobbs, 1990] J.R. Hobbs. *Literature and Cognition*. Lecture Notes. CSLI, Stanford, 1990.

[Hornby, 1974] A.S. Hornby, editor. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, 1974. revised version of 1985.

[Jespersen, 1924] O. Jespersen. *The Philosophy of Grammar*. Allen and Unwin, London, 1924.

[Jespersen, 1949] O. Jespersen. *A Modern English Grammar on Historical Principles*, volume IV – Time and Tense. Allen and Unwin, London, 1949.

[Kamp and Reyle, 1991] H. Kamp and U. Reyle. *From Discourse to Logic*. Kluwer, Dordrecht, NL, 1991.

[Kamp, 1988] H. Kamp. Discourse Representation Theory: What it is and where it ought to go. In A. Blaser, editor, *Natural Language at the Computer*. Springer Verlag, Berlin, 1988.

[Ladusaw, 1980] W. Ladusaw. *Polarity Sensitivity as Inherent Scope Relations*. Indiana University Linguistics Club, Bloomington, 1980.

[Leech, 1980] G.N. Leech. *Explorations in Semantics and Pragmatics*. John Benjamins B.V., Amsterdam, 1980.

[Leinbarger, 1980] M. Leinbarger. *Negative Polarity Items*. PhD thesis, MIT, Cambridge, 1980.

[Lundquist, 1989] L. Lundquist. Modality and text constitution. In M-E. Conte, J.S. Petöfi, and E. Sözer, editors, *Text and Discourse Connectedness. Proceedings of the Conference on Connexity and Coherence, Urbino, July 1984*. John Benjamins Publishing Co., Amsterdam, 1989.

[Magerman and Marcus, 1990] D.M. Magerman and M.P. Marcus. Parsing a natural language using mutual information statistics. In *Proceedings of the Eighth National Conference on Artificial Intelligence, AAAI-90*, Boston, MA, 1990.

[Mann and Thompson, 1986] W. Mann and S. Thompson. Relational propositions in discourse. *Discourse Processes*, 9(1):57–90, 1986.

[McDonald, 1991] D. McDonald. Reversible nlp by deriving the grammars from the knowledge base. In *Proceedings of the Workshop on Reversible Grammar in Natural Language Processing*, Berkeley, June 1991.

[Miller and Fellbaum, 1991] G. Miller and C. Fellbaum. Verbs in WordNet. *Cognition*, 1991.

[Miller and Johnson-Laird, 1976] G.A. Miller and P.N. Johnson-Laird. *Language and Perception*. Harvard University Press, Cambridge, MA, 1976.

[Moens and Steedman, 1988] M. Moens and M. Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2), 1988.

[Moravcsik, 1975] J.M. Moravcsik. Aitia as generative factor in Aristotle's philosophy. *Dialogue*, 1975.

[New, 1982] *New Scientist*, page 632, June 3rd 1982.

[Nirenburg and Defrise, 1992] S. Nirenburg and C. Defrise. Lexical and conceptual structure for knowledge-based machine translation. In J. Pustejovsky, editor, *Semantics and the Lexicon*. Kluwer, Dordrecht, NL, 1992.

[Nirenburg and Raskin, 1987] S. Nirenburg and V. Raskin. The subworld concept lexicon and the lexicon management system. *Computational Linguistics*, 13(3–4), 1987.

[Polanyi and Scha, 1984] L. Polanyi and R. Scha. A syntactic approach to discourse semantics. In *Proceedings of the International Conference on Computational Linguistics, Stanford, July 2–6, 1984*, 1984.

[Polanyi, 1987] L. Polanyi. Keeping it all straight: Interpreting narrative time in real discourse. In *Proceedings of the West Coast Conference on Formal Linguistics, Arizona 1987*, 1987.

[Porzig, 1934] W Porzig. Wesenhafte Bedeutungsbeziehungen. *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 58:70–97, 1934.

[Procter, 1978] Paul Procter, editor. *Longman Dictionary of Contemporary English*. Longman, Harlow, U.K., 1978.

[Pustejovsky, 1991] J. Pustejovsky. The Generative Lexicon. *Computational Linguistics*, 17(4), 1991.

[Pustejovsky, 1992] James Pustejovsky. Type coercion and lexical inheritance. In J. Pustejovsky, editor, *Semantics and the Lexicon*. Kluwer, Dordrecht, NL, 1992.

[Pustejovsky, 1995] J. Pustejovsky. *The Generative Lexicon: A Theory of Computational Lexical Semantics*. MIT Press, Cambridge, MA, 1995.

[Quillian, 1968] M. Ross Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*. MIT Press, Cambridge, 1968.

[Quine, 1940] W.V.O. Quine. *Mathematical Logic*. Harvard University Press, Cambridge, MA, 1940.

[Quine, 1953] W.V.O. Quine. Reference and modality. In *From a Logical Point of View*. Harvard University Press, Cambridge, MA, 1953.

[Quine, 1960] W.V.O. Quine. *Word and Object*. MIT Press, Cambridge, MA, 1960.

[Quirk *et al.*, 1985] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A Comprehensive Grammar of the English Language*. Longman, London, 1985.

[Reyle, 1986] U. Reyle. *Zeit und Aspekt bei der Verarbeitung natürlicher Sprachen*. PhD thesis, Universität Stuttgart, 1986.

[Rosch *et al.*, 1975] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 7:573–605, 1975.

[Sadock, 1974] J. Sadock. *Toward a Linguistic Theory of Speech Acts*. Academic Press, New York, 1974.

[Scha and Polanyi, 1988] R. Scha and L. Polanyi. An augmented context free grammar for discourse. In *Proceedings of the Twelfth International Conference on Computational Linguistics, Budapest, August 1988*, 1988.

[Schank and Abelson, 1977] R. Schank and R. Abelson. *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum Associates Inc., Hillsdale, New Jersey, 1977.

[Searle, 1969] J. Searle. *Speech Acts*. Cambridge University Press, Cambridge, UK, 1969.

[Sinclair, 1987] J. Sinclair. The nature of the evidence. In J. Sinclair, editor, *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins, London, 1987.

[Smadja and McKeown, 1990] F.A. Smadja and K.R. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Meeting of the Association for Computational Linguistics*, 1990.

[Sperber and Wilson, 1986] D. Sperber and D. Wilson. *Relevance – Communication and Cognition*. Harvard University Press, Cambridge, Mass, 1986.

[Taylor and Whitehill, 1981] G.B. Taylor and S.B. Whitehill. A belief representation for understanding deception. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, 24–28 August 1981*, 1981.

[Touretzky, 1986] Touretzky. *The Mathematics of Inheritance Systems*. Morgan Kaufmann, Los Altos, 1986.

[Traugott, 1989] E. C. Traugott. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language*, 65(1), 1989.

[Trier, 1931] Jost Trier. *Der deutsche Wortschatz im Sinnbezirk des Verstandes: Die Geschichte eines sprachlichen Feldes. BandI.* Heidelberg, 1931.

[Trier, 1934] J. Trier. Das sprachliche Feld. *Neue Jahrbücher für Wissenschaft und Jugendbildung*, 10:428–449, 1934.

[van Dijk and Kintsch, 1983] T.A. van Dijk and W. Kintsch. *Strategies of Discourse Comprehension.* Academic Press, Orlando, FL, 1983.

[Vendler, 1967] Zeno Vendler. *Linguistics in Philosophy.* Cornell University Press, Ithaca, 1967.

[Véronis and Ide, 1991] J. Véronis and N. Ide. An assessment of semantic information automatically extracted from machine readable dictionaries. In *Proceedings of the Fifth European Conference of the Association for Computational Linguistics*, Berlin, Germany, April 1991.

[Webber, 1988] B. Webber. Tense as discourse anaphora. *Computational Linguistics*, 14(2), 1988.

[Wiebe, 1990] J. Wiebe. Identifying subjective characters in narrative. In *Proceedings of the 13th International Conference on Computational Linguistics, Helsinki, 1990*, 1990.

[Wilks, 1975] Y.A. Wilks. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74, 1975.

[Wilks, 1978] Y.A. Wilks. Making preferences more active. *Artificial Intelligence*, 11:197–223, 1978.

[Wilks, 1986] Y.A. Wilks. Relevance and beliefs. In T. Myers, K. Brown, and B. McGonigle, editors, *Reasoning and Discourse Processes*, pages 265–289. Academic Press, New York, 1986.

[Williams, 1981] E. Williams. Argument structure and morphology. *Linguistic Review*, 1981.

[Woolf, 1974] H.B. Woolf, editor. *The Merriam-Webster Dictionary.* Pocket Books, New York, 1974.