CoreLex: Systematic Polysemy and Underspecification

A Dissertation

Presented to
The Faculty of the Graduate School of Arts and Sciences
Brandeis University
Department of Computer Science
James Pustejovsky, Advisor

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by

Paul Buitelaar

February, 1998

This dissertation, directed and approved by Paul Buitelaar's Committee, has been accepted and approved by the Graduate Faculty of Brandeis University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Dean of Arts and Sciences

Dissertation Committee

Prof. James Pustejovsky, Chair

Prof. Rick Alterman

Dr. David McDonald, Gensym Corp.

Prof. Hans Uszkoreit, Universität des Saarlandes

Prof. Dave Waltz

ACKNOWLEDGEMENTS

Perhaps, among the people that pick up this thesis, one of you is in a distant future. You browsed through the archives of the local university library and somewhere in the dust you found this copy of a thesis in early natural language processing. Probably you are a bit of a bibliophile, because in your day and age not a lot of people actually touch real books anymore. Natural language processing (NLP) has progressed onto such a level, that information can be accessed in ways that are perhaps much more convenient than by reading a book.

So, in addressing you, future reader of this prose, I would very much like to be a witness to the NLP technology that you are surrounded with. A possible future of intelligent conversation with machines has probably been an important motivation to anybody working in this field, and so it has been for me. There is also another side to this coin, perhaps an even more exciting one, which is the understanding of human language. As many others, I feel that by trying to make machines understand language, we will learn more about language ourselves. In order to do this right, we always need to try to find a balance between acquiring knowledge about language through basic research and using this knowledge in applied research. I hope this thesis is also a reflection of this basic strategy.

Now, let's move away from the future and into the past, where a number of events influenced my decision to enter this field and pursue the study of NLP by machines and by people. Although I cannot reconstruct these events entirely, I do remember I was highly fascinated by a newspaper report on machine translation. I therefore decided to focus my attention on this subject, which subsequently led to more specific research on the meaning of words, that is, on lexical semantics, mainly because after reading some books on the subject I came to the conclusion that this area should be of high importance to the general field of NLP. Also, my education at Utrecht University in the Netherlands, where I was then studying, was quite influential in this decision. Utrecht gave me a broad and liberal view on the linguistic and computational worlds, that supplied me with a wonderful basis for further explorations in the computational study of lexical semantics.

These studies eventually led me here, to Brandeis University to work with James Pustejovsky, one of the more prominent researchers in lexical semantics. As a teacher and as my advisor, James supplied me with an even broader view of the world of NLP, that gave me a new way of looking at things in semantics, involving his extensive knowledge of the general history of the field and of the philosophy of language in particular. Other professors in the department taught me what I needed to know about the complexities of computation and how to handle them. That was at times pretty hard, and I did not always pay close enough attention, but in hindsight I am very happy with what I learned. Hopefully also, the thesis got a little better by applying those skills.

In general, writing a thesis and doing the research that comes with it, is a solitary business. Fortunately, therefore, it is good to have a committee that is willing to read about and to listen to your descriptions of the work you have done so far. In this respect I would especially like to thank Hans Uszkoreit and Dave Waltz for their feedback they provided me with at different times and places. Both of them gave me a lot of direction, by making me realize which parts of my research were potentially interesting. David McDonald, unfortunately, became too occupied with his new job to supply me with much feedback, but our numerous discussions prior to the developement of this thesis have been quite significant by itself. Rick Alterman gave me some more perspective on the wider area of my research, outside of the linguistic aspects of NLP and into the realms of psychology.

Other people, who were not in my committee, gave me feedback also and I would like to thank them for that. Among these, two are prominent, both as colleagues and as friends. Marc Verhagen probably would deserve a separate paragraph, in order to thank him for all the aspects of the work reported here that have benefited from his help. I will, however, restrict myself to stressing the importance of his suggestion for me to learn the PERL programming language, which turned out to be pivotal in the whole research. Besides that, the most important thing is simply that he has always been a wonderful friend. The other person I would like to mention is Anne-Marie Mineur, because it is with her that I started this whole enterprise which eventually led to the thesis that lies before you. Although we have parted ways somewhat towards the end, our scientific paths have been quite intertwined for most of my stay here at Brandeis. And she, also, has been a good friend, although, being in Europe, geographically rather removed.

This brings me to my other friends, who played more indirect, but not less important roles in finishing this PhD enterprise. I want to thank Jeanne and Myrna for supplying me (daily) with coffee, cookies and a nice chat. Andy, Betsy, Bhuvana, Dan, David, Evelyne, Federica, Hugues, Johan, Karin, Miguel, Pablo, Piroska, Scott and Sevan, colleagues, friends and helpful visitors at Brandeis. Carol, Ellen, Jamie, Mahesh, Stephanie and Welkin for being great friends in Boston, with discussion on things that interest me sometimes more than lexical semantics... And last but not least, Clare, for playing the most prominent part of my Boston life.

Finally, I want to thank my parents, my family and friends in the Netherlands who always made that a good place to come back to and take a break from things over here. I conclude, however, by giving a separate word of thanks to Ineke, my sister, whose support (in more than one way) and advice over the years have been beyond measure and which in essence have always been crucial to me.

Waltham, Massachusetts

December 1997

ABSTRACT

CoreLex: Systematic Polysemy and Underspecification

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

by Paul Buitelaar

This thesis is concerned with a unified approach to the *systematic polysemy* and *underspecification* of nouns. Systematic polysemy – senses that are systematically related and therefore predictable over classes of lexical items – is fundamentally different from homonymy – senses that are unrelated, non-systematic and therefore not predictable. At the same time, studies in discourse analysis show that lexical items are often left underspecified for a number of related senses. Clearly, there is a correspondence between these phenomena, the investigation of which is the topic of this thesis.

Acknowledging the systematic nature of polysemy and its relation to underspecified representations, allows one to structure ontologies for lexical semantic processing more efficiently, generating more appropriate interpretations within context. In order to achieve this, one needs a thorough analysis of systematic polysemy and underspecification on a large and useful scale. The thesis establishes an ontology and semantic database (CoreLex) of 126 semantic types, covering around 40,000 nouns and defining a large number of systematic polysemous classes that are derived by a careful analysis of sense distributions in WordNet. The semantic types are underspecified representations based on *generative lexicon* theory.

The representations are used in underspecified semantic tagging, addressing two problems in traditional semantic tagging: sense enumeration (the difficulty on deciding the number of discrete senses), due to systematic polysemy; and multiple reference (NP's denoting more than one model-theoretic referent), due to underspecification. Also, traditional semantic tags that are based on discrete senses tend to be too fine-grained for practical use. For instance, WordNet has, in principle, around

60,000 different tags (synsets) for nouns alone. The CoreLex approach, on the other hand, offers a concise set of 126 tags that are inherently more coarse-grained, by taking into account systematic polysemy and underspecification.

Underspecified semantic tagging is implemented, using probabilistic classification in order to cover unknown nouns (not in CoreLex) and to identify context-specific and new interpretations. The classification algorithm is centered around the computation of a *Jaccard* (similarity) score that compares lexical items in terms of the attributes (linguistic patterns acquired from domain specific corpora) they share.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem Description

This thesis is concerned with the study of *systematic polysemy* and *underspecification*. In recent years, systematic polysemy has gained more and more interest from both theoretical and applied researchers in lexical semantics. It came to be understood that systematic polysemy – senses that are systematically related and therefore predictable over classes of lexical items – is fundamentally different from homonymy – senses that are unrelated, non-systematic and therefore not predictable. At the same time, studies in discourse analysis show that lexical items are often left underspecified for a number of related senses. Clearly, there is a correspondence between these phenomena, the investigation of which is the topic of this thesis. More specifically, the thesis deals with a unified approach to the systematic polysemy and underspecification of nouns.

Acknowledging the systematic nature of polysemy and its relation to underspecified representations, allows one to structure ontologies for lexical semantic processing more efficiently, give better explanations for non-literal use of language as in metaphor and metonymy and build more robust natural language processing systems that can generate more appropriate interpretations within context.

In order to achieve these goals one needs a thorough analysis of systematic polysemy on a large and useful scale. The thesis establishes an ontology and semantic database (CoreLex) of 126 semantic types, covering around 40,000 nouns and defining a large number of systematic polysemous noun classes that are derived by a careful analysis of sense distributions in WordNet. Some of these classes have been discussed widely in the literature, while others may not have been observed before. The semantic types that describe these classes in an intensional way are underspecified representations based on *qualia structure* as developed within *generative lexicon* theory [Pustejovsky, 1995].

The thesis further discusses the use of these representations in *underspecified* semantic tagging, addressing two problems in traditional semantic tagging that originate in systematic polysemy and underspecification: sense enumeration (the difficulty on deciding the number of discrete senses); and multiple reference (NP's denoting more than one model-theoretic referent).

Also, traditional semantic tags that are based on discrete senses tend to be too fine-grained for practical use. For instance, WordNet has, in principle, around 60,000 different tags (synsets) for nouns alone. The CoreLex approach, on the other hand, offers a concise set of 126 tags that are inherently more coarse-grained, by taking into account systematic polysemy and underspecification.

Underspecified semantic tagging is implemented, using probabilistic classification in order to cover unknown nouns (not in CoreLex) and to identify context-specific and new interpretations. The classification algorithm is centered around the computation of a similarity score that compares lexical items in terms of the attributes (linguistic patterns acquired from domain specific corpora) they share.

1.2 Thesis Outline

The central theme of the thesis is to establish and describe the use of CoreLex: an ontology for lexical semantic processing that is based on systematic polysemy and underspecification. Lexical items are assigned to systematic polysemous classes that correspond to underspecified semantic types, instead of being assigned a number of discrete senses. Each underspecified semantic type entails a number of related interpretations that are to be generated within context. Figure 1.1 gives an overview of the ideas behind CoreLex as they are being employed in this thesis.

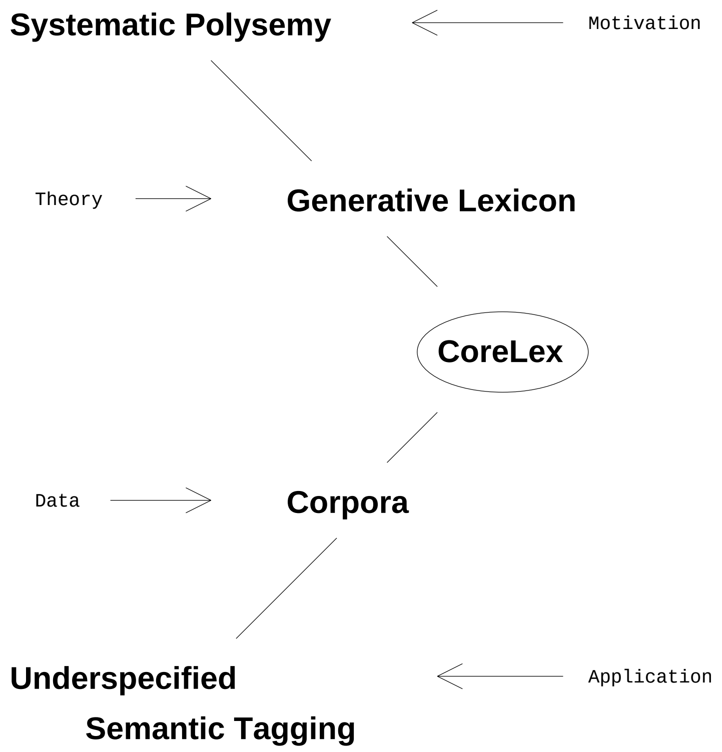

Figure 1.1: Overview of the thesis

A unified approach to systematic polysemy and underspecification pulls together both the motivation for this research, which originates in the study of systematically related senses in lexical semantics, and the application of these studies in tagging lexical items for discourse processing with underspecified representations that reflect their related senses. This approach further employs a theoretical underpinning for these lexical semantic studies (*generative lexicon* theory [Pustejovsky, 1995]) and ties this in with an empirical approach to semantic tagging, using various corpora to derive relevant underspecified lexical semantic representations.

At the core of the thesis are the ideas of systematic polysemy and underspecification. Systematic polysemy can be distinguished from homonymy by looking at the behavior of lexical meaning in discourse. Homonys, like 'bank' will be referred to in a coherent discourse in only one of their (unrelated) senses. This, in fact, has been established empirically [Gale *et al.*, 1992]. Systematically polysemous nouns, on the other hand, will occur with different, but related interpretations, throughout a coherent discourse. The noun 'school', for instance, can refer in the same discourse to a *group of people, a* *learning process, an institution* and *a building*[1] [Bierwisch, 1982]. This noun will thus be underspecified for any one of these related senses, unless in a local part of the discourse (a coherent discourse segment, a sentence, or even part of a sentence – a collocation) one sense is more prominent than any of the others. Still, even then, these other senses will be present, although on the 'background'. We can illustrate this with example (1) from the BROWN corpus:

(1) [A long book heavily weighted with military technicalities]$_{NP:event-object-content}$, in this edition it$_{PRON:event}$ is neither so long nor so technical as it$_{PRON:content}$ was originally.

The head of the main NP in this sentence, 'book', is underspecified for the interpretations: **event, physical_object** and **informational_content** corresponding to some of the systematically related senses of this noun. Still, if we look at the two pronouns in the same sentence that refer back to that same NP, we find that each is specified for one of these senses. We thus need an underspecified representation of the lexical semantics of the noun 'book', in order to accommodate both of the anaphoric references.

Underspecified representations, based on systematic polysemy, can be acquired on a large scale by carefully analysing existing lexical semantic resources like WordNet. The thesis takes the view that underspecified lexical semantic representations are already there, although only in implicit ways. In order to make these representations explicit, a number of steps have to be taken that will highlight related senses that occurring with classes of nouns and can therefore be said to be systematic. Figure 1.2 shows the three derivation steps that were needed to acquire CoreLex from WordNet.

The derivation reduces the senses for nouns in the WordNet database to a set of basic types, corresponding to the (sub-)top levels of the WordNet hierarchy and then groups the nouns together in systematic polysemous classes according to their basic type distributions. Additionally, the systematic polysemous classes are represented through underspecified semantic type definitions that take the form of a *qualia structure*, as adopted from generative lexicon theory. Qualia structure represents several relations between nouns and other (types of) lexical items, covering inheritance (`Formal`), modification (`Const`) and thematic role assignment (`Agt-Telc`).

[1] These same systematically polysemous senses apply also to the homonymic "money_institution" sense of 'bank'.

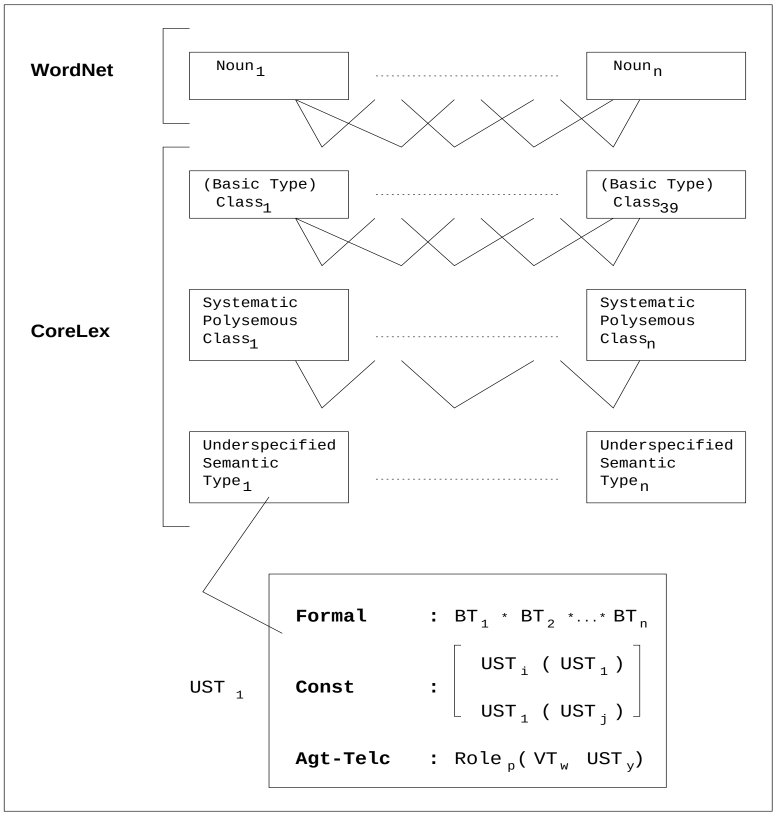

Figure 1.2: Derivation steps to acquire CoreLex from WordNet

Of these three sources of knowledge (inheritance, modification and thematic role assignment), only knowledge on inheritance is supplied by WordNet. Knowledge on modification and thematic role assignment has to be acquired from other sources, notably domain specific corpora, also because such knowledge varies widely between domains. This brings us to a different aspect of the thesis, dealing with knowledge acquisition from text.

Here, the viewpoint is taken that we can analyse lexical semantic structure by abstraction over the functional structure of sentences and the lexical items that occur in these. The qualia structure for a particular type of nouns is then the accumulation of the functional structures in which nouns of this type occur, relative to the particular corpus and domain that we are looking at. The functional structure of sentences, in turn, can be analysed by taking into account a combination of syntactic (phrase) structure and the semantic type of lexical items. For instance, the sentence (2a) with phrase structure (2b) gives a functional structure like that of (2c).

(2a) Stocks collapsed in early afternoon

(2b) Stocks_{NP} collapsed_{VG} $[\text{in}_{PREP}$ $[\text{early}_{ADJ}$ $\text{afternoon}_{NOUN}]_{NP}]_{PP}$

(2c) $\text{Stocks}_{NP-agent}$ collapsed_{VG} $[\text{in}_{P}$ $[\text{early}_{A-mod}$ $\text{afternoon}_{N-head}]_{NP-time}]_{PP}$

The functional structure (2c) then results in the following 'contributions' to the qualia structure of 'stocks' and 'afternoon'. (Qualia structures derived from complete corpora, representing particular domains, will have type information that abstracts over lexical instances like these. For instance, instead of the verb 'collapse', the **agent** and **time** roles in the `Agt-Telc` will be filled by a (number of) verb *type(s)*.)

'stocks'

`Agt-Telc` : **agent**(collapse, stocks)

'afternoon'

`Const` : **modification**(early, afternoon)

`Agt-Telc` : **time**(collapse, afternoon)

The lexical knowledge that is acquired in this way 'spells' out the qualia structure or *underspecified representation* for each (type of) systematically polysemous noun. The next step in the research is to tag (perhaps *annotate* would be a better term) corpora with this acquired knowledge, in order to facilitate the correct processing of underspecified discourse referents as discussed above. This approach is therefore called *underspecified semantic tagging*, addressing the following two problems in traditional semantic tagging that originate in systematic polysemy and underspecification. Systematic polysemy leads to non-discrete senses (or rather interpretations) that cannot be easily enumerated and are better represented through an underspecified representation. At the same time, underspecified discourse referents have multiple referents (as in model-theory) that cannot be handled by discrete sense-based, single referent semantic tags.

Also, such traditional semantic tags that are based on discrete senses and sense enumeration tend to be too fine-grained for practical use. In WordNet (the most widely used database for semantic tagging) there are, in principle, around 60,000 different tags (sense level synsets) for nouns alone. The CoreLex approach on the other hand offers a concise set of 126 tags that are inherently more coarse-grained, by taking into account systematic polysemy and underspecification.

Underspecified semantic tagging with this set of tags is implemented by using the CoreLex database, the acquired underspecified representations and a system of probabilistic classification in order to cover unknown nouns (those nouns that are not in the CoreLex database) and to identify context-specific and new interpretations. The classification algorithm is centered around the computation of a similarity score (*Jaccard* [Grefenstette, 1994]) that compares objects (nouns and classes of nouns) in terms of the attributes they share. Here, attributes are linguistic patterns (functional relations) as they appear in the acquired underspecified representations.

Summarizing all of this, the thesis deals with the construction and use of CoreLex as a lexical semantic database, a top-level ontology and a coarse-grained semantic tag set. None of these are mutual exclusive. In fact, they are just different facets of an integrated approach to lexical semantic analysis in language understanding.

1.3 The CoreLex project

1.3.1 About CoreLex

Some history

Many of the ideas described in this thesis are based on previous research within the framework of *generative lexicon* theory [Pustejovsky, 1991a] [Pustejovsky, 1995]. The idea for CoreLex itself originates out of an NSF-ARPA funded research project on the Core Lexical Engine, a joint research project of Brandeis University and Apple Computers Inc. [Pustejovsky *et al.*, 1995]. Results of this research have been published in various theoretical and applied papers [Pustejovsky, 1994b] [Johnston *et al.*, 1995] [Pustejovsky, 1996]. The research described in this thesis however is the first comprehensive attempt to actually construct an ontology according to some of the ideas that arose out of this accumulated research and investigate its use in both classification and semantic tagging.

About the name

The name of the project points to some of the basic ideas behind it. CoreLex describes the *core* of lexical semantic structure from which more peripheral interpretations can be generated, but it also describes the *core* of the vocabulary of a language. General inference mechanisms of reasoning with (lexical semantic) knowledge extend the former with context dependent interpretations that are derived within discourse processing. The latter is extended with machine learning techniques that add domain specific lexical items and classify them in a probabilistical fashion. The thesis deals with the theory and implementation of both these extensions to CoreLex, using the same pattern matching and classification techniques in both cases.

1.3.2 CoreLex on-line

The thesis is accompanied by a *web page* on the CoreLex project:

`http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html`

The page includes a detailed presentation of the structure and content of CoreLex as it results from this research. The ontology consists of 126 underspecified semantic types, corresponding to 324 systematic polysemous classes that were derived from WordNet. Each of these classes is presented, complete with all their instances. Figure 1.3 gives a snapshot of this overview. It can be found at:

`http://www.cs.brandeis.edu/~paulb/CoreLex/overview.html`

1.4 Thesis Overview

Chapter 2 Explains some of the historic and contemporary background to the idea of systematic polysemy. I discuss some early observations that were made first by Weinreich and somewhat later by Apresjan, distinguishing polysemy from homonymy. This is followed by a discussion of two opposite approaches towards systematic polysemy, one that takes it to arise from lexical semantic structure, while the other sees it as a much less structured phenomenon that is governed by pragmatic constraints. Finally, we reach into computational research that tries to deal with systematic polysemy

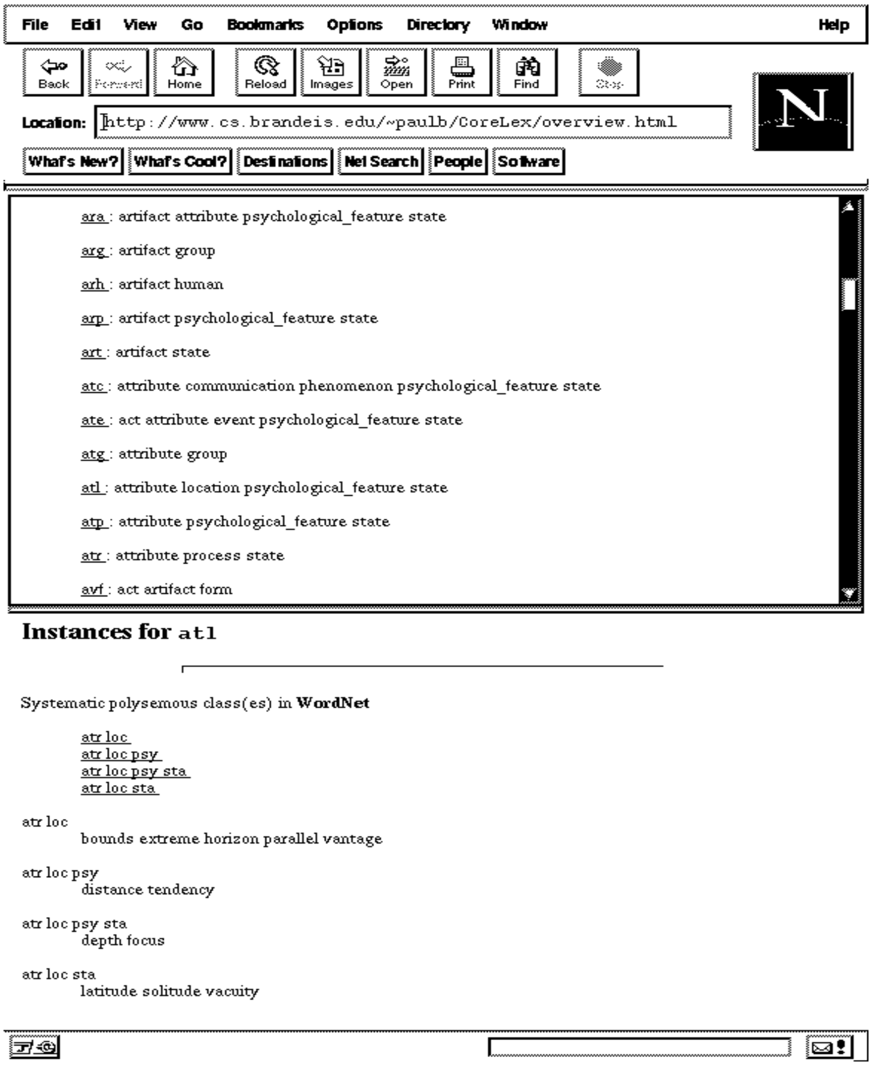

Figure 1.3: Snapshot of CoreLex overview

and adjacent phenomena (metonymy, metaphor, sense extension, coercion, etc.) by means of rules, representations and inference mechanisms.

Chapter 3 Discusses Pustejovsky's proposal for a *generative lexicon*, which takes systematic polysemy central to the design of a semantic lexicon. Generative lexicon ($\mathcal{GL}$) theory incorporates a lot of the ideas that were discussed in Chapter 2 and integrates them into a coherent approach to systematic polysemy and its consequences for lexical and compositional semantics. The chapter describes $\mathcal{GL}$ both in detail as it emerges from existing literature and how it will be used in the thesis.

Chapter 4 Describes how to construct an ontology (CoreLex) that is based on $\mathcal{GL}$ theory using existing resources like WordNet. This starts by comparing the different lexical items in that database in terms of their polysemous bahaviour. Those with the same polysemous patterns are grouped into *systematic polysemous classes*, which in turn are described by socalled *underspecified semantic types* along the lines of $\mathcal{GL}$ theory.

Chapter 5 Describes how to automatically acquire underspecified lexical semantic (qualia structure) representations from corpora. The chapter discusses in detail the finite state methods for lexical knowledge acquisition, starting with tokenizing of the text, part-of-speech tagging and stemming, through a shallow syntactic analysis that extracts verb groups and nominal phrases, unto some forms of semantic analysis that include thematic role assignment and part-whole analysis. Also, a method for extracting default types in lexical semantic structure is discussed.

Chapter 6 The chapter discusses two major problems for traditional semantic tagging based on word sense disambiguation, *multiple reference* in discourse and *sense enumeration.* Although both of these are not a problem in working with homonyms, on which most of the work in semantic tagging has focused, they are in capturing lexical items that are polysemous. Besides this theory behind semantic tagging with underspecified representations, the chapter also presents some experimental results of an actual implementation. The semantic tagger uses a classification algorithm, at the core of which is a similarity measure (*Jaccard*) that compares objects (lexical items) in terms of the attributes (linguistic contexts) they share.

Chapter 7 Conclusions and an outlook on further research.

Appendix A Overview of all 126 underspecified semantic types in CoreLex and the *basic* types they include.

Appendix B Overview of input/output for the programs that were developed in the context of this thesis.

Chapter 2

Systematic Polysemy

2.1 Introduction

Polysemy is either of a *contrastive* or a *complementary* nature [Weinreich, 1964]. Contrastive senses (or *homonyms*) are unrelated to each other, for instance:

(3a) We walked along the bank of the Charles river.

(3b) Did he have an account at the HBU bank?

Complementary senses on the other hand do not contrast each other, but seem to be related in a systematic way. For instance door has the two related senses of an **aperture** to walk through (4a) and something **physical** that can be painted (4b) :

(4a) He walked through the door on the left.

(4b) He painted every door in the house.

The systematic relation between complementary senses has been investigated by a number of researchers. The following section discusses some early contributions by [Weinreich, 1964], [Weinreich, 196x] and [Apresjan, 1973]. Weinreich makes a number of interesting observations about the nature of polysemy and some of the systematicity behind it, but falls short of giving particular examples, apart for illustration. Apresjan on the other hand presents a comprehensive list of examples for verbs, nouns and adjectives in Russian, besides giving some useful definitions. The next section takes a close look on a fundamental difference in analysis of systematic polysemy in terms of semantics and pragmatics. Systematic polysemy either arises from lexical semantic structure or it is a much less structured phenomenon that is governed by pragmatic constraints. The chapter closes with a presentation of a number of computational approaches that deal with systematic polysemy and related phenomena (metonymy, metaphor, sense extension, etc.). They are grouped into four sections, the first two of which are more AI inspired (*conceptual graphs* and *abduction*) whereas the second and third are more linguistically oriented (*lexical rules* and *lexical representations*).

2.2 Early observations

Our discussion of previous work in this field starts with [Weinreich, 1964], [Weinreich, 196x] and [Apresjan, 1973], neglecting much of the work that preceded them, including for instance [Stern, 1931] and other research on linguistic semantics that predates current approaches to linguistic theory and was still largely grounded in etymological studies. Such an approach to lexical semantics is no longer very common and the emphasis switched to a synchronic study of meaning and meaning change. However, it should be noted that change of meaning is obviously a diachronic phenomenon. Therefore perhaps, a refocus on diachronic semantics is asked for, although in close interaction with synchronic studies. This is well beyond the scope of the thesis and we therefore concentrate on synchronic studies here, starting with Weinreich and Apresjan.

2.2.1 Weinreich

Lexicographers have always treated lexical meaning as something discrete that can be quantified in a certain number of distinct 'senses'. [Weinreich, 1964] criticizes this approach and, in discussing Webster's third edition, questions the scientific value of such arbitrary groupings of senses: '... we find the meaning of the verb [turn] classified into 115 ultimate sub- or subsubsenses. What validity does this grouping have? ... Consider the sentence: "This is the man who turned the figurines." The ambiguity is clear-cut: we cannot tell whether the man rotated the figurines or shaped them on a lathe. The ambiguity therefore establishes at least two senses for the transitive verb *to turn*. On the other hand, the different senses of turn in *he turned the crank* and *he turned a somersault* may safely be attributed to the very different semantic types of the word *crank* (a name of a concrete physical object) and *somersault* (a name of an act characterized by rotary motion). ... Moreover, we must bear in mind that the ambiguity of a polysemous word is usually restricted to certain contexts; thus, in the cited example, the ambiguity of *turn* is perceived when the direct object is *figurine*, but it does not materialize when the object is *somersault*, since somersaults are not expected to be shaped on a lathe.'

Weinreich then proceeds in pointing out an interesting analogy with 'context-bound ambiguity in phonology, where we have learned to cope with it under the rubric of neutralization. In semantics the problem is far more intricate: the number of basic elements is very large and the contexts are difficult to classify.' In Weinreich's view '...the dictionary shrinks from a specification of the relevant class of subject nouns and heaps examples upon examples'. Instead he argues for an 'approriate metalanguage' with which to describe classes of examples. 'Take the citation sentence, "[The] discussion turned ... on the feasibility of the scheme." ... Now clearly this subsense of *turn* is evoked by the noun which is here the subject; the grammatically similar sentences, *The bull turned on the matador* and *The visitor turned on his heel* are obviously different instances. ... the relevant sense of *turn* seems to be evoked uniquely if the subject is a noun like *story, book, plot, dream, play, thought, opinion*. These are all names of things which can be *about* something...'

In [Weinreich, 196x] on the 'Dictionary of the Future' it is further noted that '[o]bviously the sign is not always used by people - even by one and the same person - to convey *precisely* the same meaning as on all other occasions. ... But the same sign, recurring, does convey *some* of the same meaning. Otherwise communication would be impossible ... It is this constant, institutionalized aspect of the meaning of signs that descriptive semantics must deal with, without denying the existence of a noninstitutionalized margin of meaning. To the extent that a speaker puts idiosyncratic meanings into a sign, or that a hearer finds idiosyncratic meanings in it, they are not suitable sources for the semantic description of the language. ... Only the learner -the child, the foreigner- in noticing a discrepancy between a feature of the denotatum and the significatum, will adjust his significatum; the normal, mature member of the speech community will instead detect a joke, a metaphor, a lie, or a lapse. Only when such discrepancies become frequent and systematic will an adjustment in the common significatum take place: this is the adult's way of keeping up with semantic change.'

2.2.2 Apresjan

[Apresjan, 1973][1] agrees with previous researchers like Weinreich 'that the usual explanatory dictionaries greatly exaggerate the measure of discreteness of meanings and are inclined to set clear-cut borders where a closer examination of the actually existing variety of texts reveals only a vague intermediate area of overlapping meanings.' In view of this he distinguishes three such cases of polysemy:

- 'Of particular interest is that instance of ambiguity of the type under consideration, which manifests itself in distinctions only on the level of signification, but not on the level of denotation, or reference. ... the effect of the oscillation of meanings between "action" and "place" appears not because there is an intermediate area between them, but because their referents coincide.'

- 'The ambiguity of the given speech segment is intentional and is used as the basis for a pun, a poetic metaphor, or another stylistic device based on the simultaneous use of several meanings of a word ... [but] the esssence of the pun, the poetic metaphor, and similar stylistic devices lies precisely in the fact that attention is focussed at once on two clearly isolated meanings ...'

- 'A given speech segment contains an instance of the syncretic manifestation of two different meanings of a polysemantic word. By way of illustration we shall examine two meanings of the verb ... [to battle as in:] "The inquisiton battled heresy" ... [T]he corresponding phrases can be and often are comprehended both in the sense that the subject of the struggle ceases the activity of the heretics and in the sense that he achieves physical destruction of them as means of eradicating heresy.'

To distinguish polysemy from homonyms Apresjan accepts that 'the existence in a given language of at least a pair of words with a given combination of meanings is considered to be an indisputable

[1] Originally published in Russian in 1971

sign of polysemy ... If the given combination is unique, that is if it occurs only once in the whole vocabulary of the langugae, then we are dealing not with a polysemantic word but with homonyms. ... we can speak of different types of polysemy from the point of view of their relative remoteness from homonymy. Closest to homonymy are some types of metaphorically motivated polysemy ... Metonymically and functionally motivated polysemy is, generally speaking, removed a step further from homonymy ...'

Within polysemy there is a further distinction between *regular* and *irregular* polysemy:

> 'Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language, there exists at least one other word B with the meanings b_i b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and if a_i and b_i, a_j and b_j are nonsynonymous.'
>
> 'Polysemy is called irregular if the semantic distinction between a_i and a_j is not exemplified in any other word of the given language'

It is an interesting observation that 'regularity is a distinctive feature of metonymical transfers' while 'irregular polysemy is more typical of metaphorical transfers'.

Finally, Apresjan notes that '[i]n addition to metonymical transfers, several other processes give rise to regular polysemy. Among them are semantic analogy ... compression of phrases ... [and] various word-formation processes ... Regular polysemy is similar to word formation also in the sense that many of its types are productive; in other words, for any word which has a meaning of type "A", it is true that it can be used in a meaning of type "B" as well ...'

2.3 The semantics and pragmatics of systematic polysemy

Systematic polysemy has some interesting consequences for a theory about the relation between language and the world. As noted by both Weinreich and Apresjan polysemous lexical items have more than one referent that can be expected to relate to each other in some way. There has been some discussion in the literature about what the nature of this relation exactly is: *semantic* [Pustejovsky, 1991a], *conceptual* [Bierwisch, 1982] or *pragmatic* [Nunberg, 1979] [Ruhl, 1989] [Mercer, 1993] [Lascarides and Copestake, 1995]. Bierwisch' conceptual relation can be summarized as 'semantic structure plus (linguistic) context' and I therefore group it together with the semantic relation, at least for the purposes of this thesis[2].

The semantic analysis takes the relation to be within language, between lexical items or between different semantic aspects of the same lexical item. A pragmatic analysis rather sees a relation between objects in the world. The two approaches can be represented schematically as in Figure 2.1:

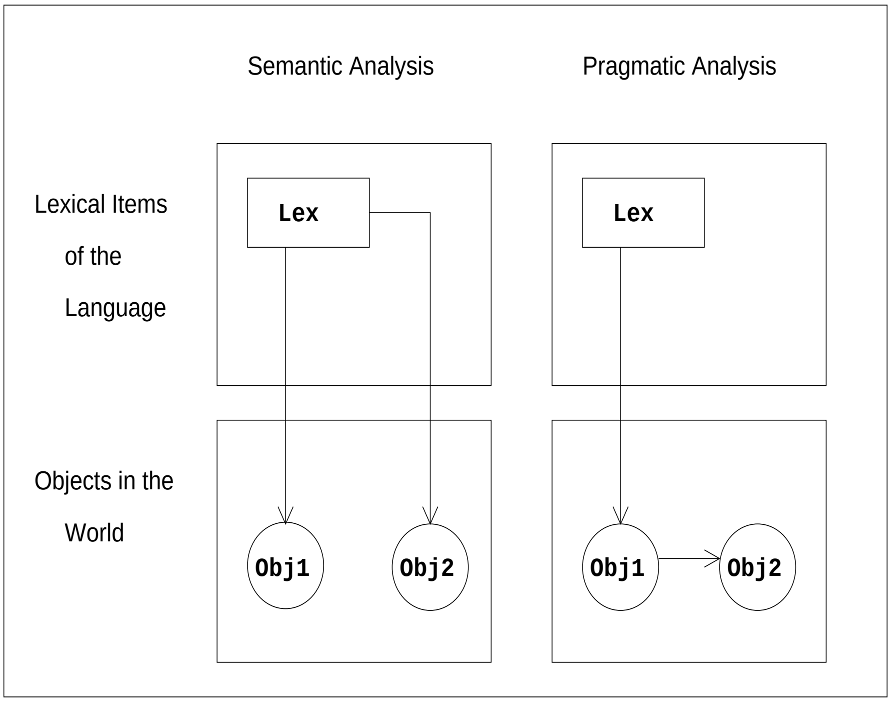

Figure 2.1: Two analyses of systematic polysemy

In a semantic analysis a lexical item *Lex* points to more than one object at the same time. In a pragmatic analysis *Lex* points to object Obj_1 and points indirectly to object Obj_2, because of the relation between objects Obj_1 and Obj_2.

[2]A conceptual relation assumes a conceptual *level*, which is hard to identify. Nevertheless, such a conceptual level inbetween semantics and pragmatics could help to 'liberate' lexical semantics from a formal semantic harness, that stresses in particular compositionality and looses sight of the wider associations that most lexical items have.

For instance, examples (5a) through (5d) illustrate the systematic polysemy of school [Bierwisch, 1982].

(5a) *a group of people*:

The school went for an outing.

(5b) *a learning process*:

School starts at 8.30

(5c) *an institution*:

The school was founded in 1910.

(5d) *a building*:

The school has a new roof.

From a semantic point of view, this noun points to more than one object at the same time. Alternatively, a pragmatic analysis would argue that school points to a certain object in the world, which in turn would have some relation with other objects in the world, independent from the *word* school. Or as [Nunberg, 1979] puts it: '... we will presume that we have a way of getting from a name to its designatum, or from a demonstrative to a physically present demonstratum, and proceed to ask how we get from that thing to something else. So it will be distinct from a theory of linguistic meaning, of how we get from words to things; it will be concerned exclusively with speakers and things, and will have nothing to say about words at all.'

Although a distinction in semantic or pragmatic is not crucial for a representation of systematic polysemy it is still an important question in respect to how much knowledge should be represented in the lexical entry. A semantic approach would store as much as possible, because all inferences would be made from within the lexical representation. According to a pragmatic analysis the entry would be down to basics and the inferences would be made on the basis of global principles or 'conventions' [Nunberg, 1979].

2.4 Computational approaches

2.4.1 Conceptual Graphs

[Sowa, 1993] cites an example from [Kayser, 1988] to illustrate how metonymies can be handled by *conceptual graphs*, a graphical knowledge representation language based on semantic networks. Kayser's example is the 'Prix Goncourt', a prestiguous French literary prize, for which he found the following seven related senses:

Prize	The Prix Goncourt was awarded to X.
Money	X turned over his Prix Goncourt to the Red Cross.
Person	The Prix Goncourt was congratulated by the President.
Panel	The Prix Goncourt admitted a new judge.
Book	Could you go buy the Prix Goncourt for me at bookstore X?
Time	Since his Prix Goncourt, he has become arrogant.
Institution	The Prix Goncourt perverts the literary life.

With the conceptual graph in Figure 2.2 as background knowledge, each metonym of Prix Goncourt can be determined by following a path in the graph. Three sample inferences are (from [Sowa, 1993]):

- Prize: The verb awarded in the input maps to the concept [AWARD]. The corresponding concept in the background graph is linked to [PRIZE] by the relation (PTNT). A maximal join of the input graph to the background graph starting with the two concepts of type AWARD would automatically associate PG with the node [PRIZE].

- Book: The verb buys maps to [BUY], which would prefer a nonhuman physical entity as patient. Buying a prize is possible, but that might suggest bribing the panel. The background knowlege that bookstores sell books would give a strong preference for BOOK, which would unify with [LITERARY-WORK] (although this is another example of metonymy, since book could refer to the literary work or to a physical object in which the work is printed).

- Person: The verb congratulates maps to [CONGRATULATE], with its selectional constrains for PERSON or a subtype such as AUTHOR. Judges are also persons, but the node [JUDGE:*] is marked as plural by the symbol * and is therefore unlikely to be indicated as *the* Prix Goncourt. Information about salience should also be marked on the graph: [AUTHOR], [PRIZE] and [LITERARY-WORK] are more salient and hence more likely to be selected.

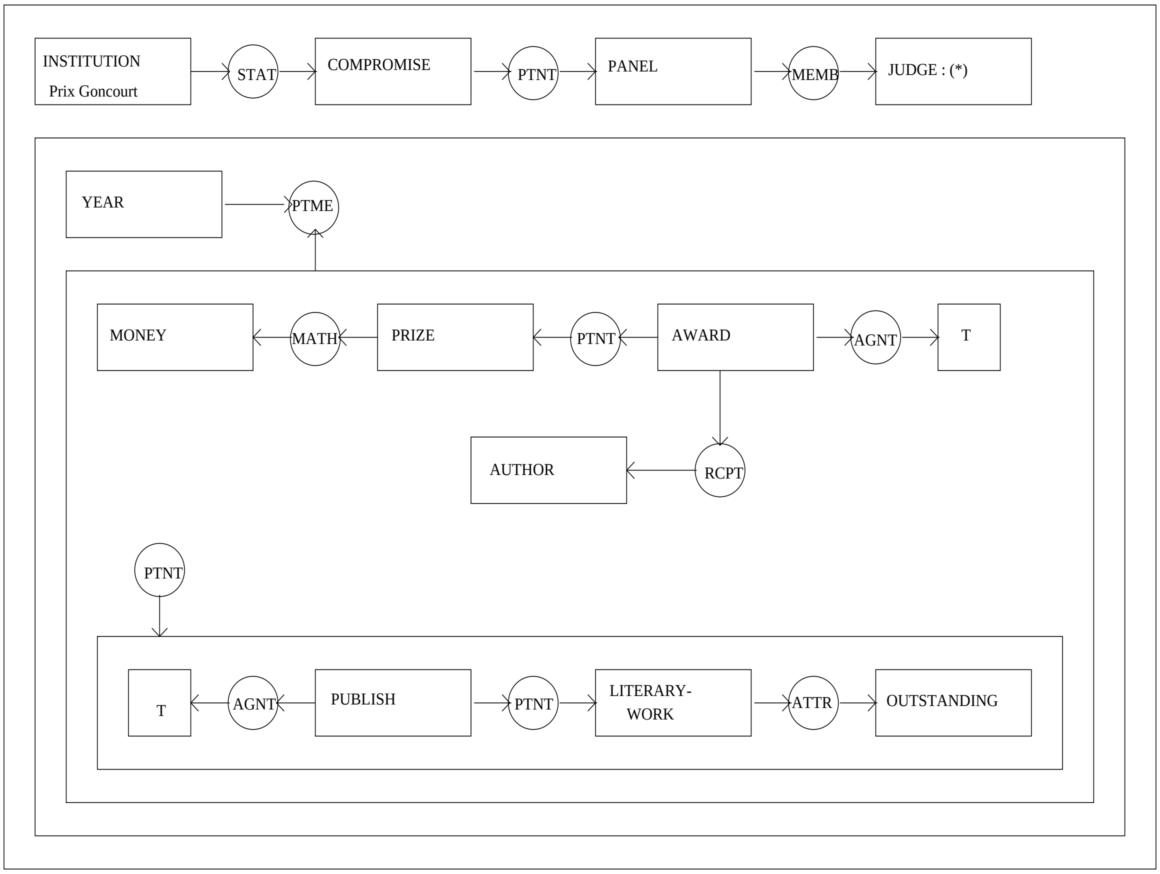

Figure 2.2: Conceptual graph for Prix Goncourt

2.4.2 Abduction

Conceptual graphs are based on semantic networks, but designed specifically to handle linguistic constraints, especially selection restrictions on thematic roles. In contrast, the knowledge representation schemes that [Hobbs *et al.*, 1993] employs to deal with similar cases are based on standard first-order logic without any specific linguistic considerations. The emphasis is rather on the general inference technique of *abduction*, which can probably best be paraphrased by 'reasoning backwards', by assuming something to be true try to find a reason for it to be true.

For example, sentence (6a) can only be understood if one assumes something like sentence (6b), which corresponds to the logical form (6c):

(6a) The Boston office called

(6b) Some person at the Boston office (who called).

(6c) $(\exists, y, z, e)\ call(e, x) \wedge person(x) \wedge rel(x, y) \wedge office(y) \wedge Boston(z) \wedge nn(z, y)$

Now '[w]e must prove abductively the logical form together with the constraint "call" imposes on its agent, allowing for a coercion. ... The sentence can be interpreted with respect to a knowledge base of mutual knowledge that contains the following facts:'

$Boston(B_1)$

$office(O_1) \wedge in(O_1, B_1)$

$person(J_1)$

$work\_for(J_1, O_1)$

$(ALL\ y, z)\ in(y, z) \supset nn(z, y)$

$(ALL\ x, y)\ work\_for(x, y) \supset rel(x, y)$

The proof results in a 'proof graph' that can be paraphrased as follows:

$person(x)\ :\ person(J_1)$

$office(y)\ :\ office(O_1)$

$Boston(z)\ :\ Boston(B_1)$

$work\_for(x, y) \supset rel(x, y)\ :\ work\_for(J_1, O_1)$

$in(y, z) \supset nn(z, y)\ :\ in(O_1, B_1)$

$\boxed{\text{call'(e,x)}}\ :\ call'(e, J_1)$

According to [Hobbs *et al.*, 1993] 'three local pragmatics problems have been solved as a by-product. We have resolved "the Boston office" to O_1". We have determined the implicit relation in the compound nominal to be *in*. And we have expanded the metonymy to "John, who works for the Boston office, called."'

Notice the use of the term 'local pragmatics'. This puts the approach clearly on the side of Nunberg in the semantics/pragmatics debate, which is in line with the idea of a general inferencing approach that is based on *commonsense knowledge*, rather than assuming a linguistic or lexical semantic level of knowledge. Next to the abduction approach there are other and more direct implementations of Nunberg's program as reported in [Dölling, 1994] and [Verschuur, 1993], both of which use higher order logics (λ calculus).

2.4.3 Lexical Rules

Opposite to the ideas of general knowledge representation schemes (*conceptual graphs*) or general inference schemes (*abduction*) is the approach of stipulating a number of lexical rules, that can be applied to specific classes of lexical items in order to change their meaning in some way[3].

Lexical Operations

[Copestake and Briscoe, 1993] propose socalled *lexical operations* in the context of unification-based grammars. A lexical operation is a feature structure that takes a feature structure as input and outputs another feature structure[4]. This is illustrated with the example of *animal_grinding* in Figure 2.3 that turns an animal into the food it produces. It also turns a **count-noun** into a **mass-noun**.

$$\left[\begin{array}{l} \textbf{animal\_grinding} \supset \textbf{lexical\_rule} \\ \left[\begin{array}{l} \textbf{count-noun} \\ \text{ORTH} = \boxed{1} \\ \text{RQS} = \left[\begin{array}{l} \textbf{animal} \supset \textbf{ind\_obj} \\ \text{EDIBLE} = + \end{array}\right] \end{array}\right] \Rightarrow \left[\begin{array}{l} \textbf{mass-noun} \\ \text{ORTH} = \boxed{1} \\ \text{RQS} = \textbf{food\_substance} \supset \textbf{substance} \end{array}\right] \end{array}\right]$$

Figure 2.3: Lexical operation for *animal_grinding*

[3]Lexical rules as described here are not like those of LFG [Bresnan, 1982] or HPSG [Pollard and Sag, 1994], which are mostly about mapping syntactic onto semantic stucture.

[4]Both feature structures represent *lexical signs* as familiar from HPSG [Pollard and Sag, 1994].

For example, we can apply this rule to the feature structure that corresponds to the lexical item haddock[5]. as given in Figure 2.4: the input for the rule is an 'individuated object that is edible'; the output is a 'food substance which purpose (TELIC) it is to be eaten'.

$$\left[\begin{array}{l} \textbf{count-noun} \\ \textsc{orth} = \textbf{haddock} \\ \textsc{syntax} = \left[\textsc{count} = +\right] \\ \textsc{rqs} = \left[\begin{array}{l} \textbf{animal} \\ \textsc{sex} = \textbf{gender} \\ \textsc{age} = \textbf{scalar} \\ \textsc{edible} = \textbf{boolean} \\ \textsc{physical\_state} = \textbf{solid} \\ \textsc{form} = \textbf{individuated} \end{array}\right] \end{array}\right] \Rightarrow \left[\begin{array}{l} \textbf{mass-noun} \\ \textsc{orth} = \textbf{haddock} \\ \textsc{syntax} = \left[\textsc{count} = \text{-}\right] \\ \textsc{rqs} = \left[\begin{array}{l} \textbf{food\_substance} \\ \textsc{telic} = \textbf{eat} \end{array}\right] \end{array}\right]$$

Figure 2.4: Application of *animal_grinding* to haddock

[Copestake and Briscoe, 1995] extends on the idea of unification-based lexical rules for handling systematic polysemy. To deal with the semi-productivity of lexical rules they introduce a probabilistic constraint[6] on the application order of the rules. Figure 2.5 gives an example for rabbit with the lexical rules that are associated with this lexical item.

$$\left[\begin{array}{l} \textbf{lex-count-noun} \\ \textsc{orth} = \textbf{rabbit} \\ \textsc{cat} = \boxed{\text{noun-cat}} \\ \textsc{sem} = \boxed{\text{obj-noun-formula}} \\ \textsc{prob} = \textbf{0.4} \\ \textsc{lrs} = \\ \quad grinding(0.05), meat\_grdg(0.3), \\ \quad meat\_grdg + portioning(0.15), fur/skin\_grdg(0.1) \end{array}\right]$$

Figure 2.5: Lexical entry for rabbit with probabilistic lexical rules

The order in probabilities for the lexical rules in LRS corresponds to the differences in interpretation

[5] RQS stands for *relativized qualia structure*, see Chapter 3 and [Calzolari, 1991].

[6] See also [Sanfilippo, 1993].

in sentences (7a) through (7c). According to [Copestake and Briscoe, 1995] the interpretation of rabbit as *rabbit meat* in (7a) is 'neutral', whereas the interpretations in (7b) and (7c) are more on the periphery.

(7a) John prefers rabbit.

(7b) John wants three rabbits.

(7c) The diners ordered three rabbits.

Lexical Rules

[Viegas *et al.*, 1996] employ similar lexical rules for large scale lexicon acquisition. The number of rules and the scale of their application leads to questions about their performance. When should lexical rules apply and how should they be represented for an optimal performance?

There are three options for the time of application of lexical rules:

- Acquisition time
- Lexicon load time
- Run time

All three of the options have drawbacks. If the rules apply at acquisition time human intervention can check the output, but may easily be flooded by the number of generated lexical entries. At lexicon load time no manual intervention is assumed, but overgeneration will cause a problem. At run time there are similar advantages and disadvantages, on top of which the index of head forms to be used by syntactic and semantic processes is no longer completely known. The best option seems to be a hybrid approach with rules applying at acquisition time to produce new entries, but also at lexicon load or run time to deal with unknown lexical items or interpretations. More studies need to be made however to determine the best time and combinations of rule application.

The representation of lexical rules can either be *global* or *local*, that is: the lexical rules can be specified locally in each entry, or globally for all entries with additional typing information to restrict their application to certain classes of entries. A local approach results in a representation of rules as in the rabbit example above. Setting up such representations however requires 'tedious work at lexicon acquisition time', something which can be avoided with a global approach. The drawback of that option however is the possibility of overgeneration if typing restrictions are not strict enough.

Lexical Implication Rules

An interesting collection of lexical rules is introduced in [Ostler and Atkins, 1993]. It comprises of a database of about 130 'types of systematic meaning shift' that are called Lexical Implication Rules (LIR's). It draws on previous work in Fillmore, [Apresjan, 1973], Leech, Levin, Miller, Clark and Clark, Aronoff, Lehrer, Talmy, Levi and [Pustejovsky, 1991a]. Each LIR is documented with 'separate lists (as complete as possible) of the sets of words which participate in each type, and those which might be expected to, but do not'.

Some examples of LIR's are:

- LIR **Container – Amount/Contents**

  NC singular: container (purpose-built only) → NC: amount it contains/its contents

  (8a) The glass broke.

  (8b) Add a glass of wine.

  (8c) Don't drink the whole glass.

  Also: trunk, tank, jug, hamper, bucket, basket,...

- LIR **Constitution – State**

  NU: form of government → NC: state with that form

  (9a) Democracy is at risk.

  (9b) Democracies must take action.

  Also: oligarchy, monarchy, tyranny, dictatorship,...

Some LIR's are between different syntactic categories or involve morphological processing, which make them not purely about systematic polysemy in (lexical) *semantics*:

- LIR **Vehicle Verb**

  NC singular: vehicle → VTI (verb trans/intrans): to travel/transport using that vehicle

  (10a) He's got a new cycle.

  (10b) Let's cycle into town.

  Also: motor, ferry, canoe, ship, jet,...

- LIR **'-ful' Derived Noun**

  NC singular: container → $[\text{NC+ful}]_{NC}$: amount it contains/its contents

  (11a) I don't use a thimble.

  (11b) Add a mere thimbleful of brandy.

  Also: trunk, tank, jug, hamper, bucket, basket,... (purpose-built containers) and spade, spoon, fork, hand, fist belly,... (other items that can be viewed as containers)

The implementation of a LIR is not discussed in much detail in [Ostler and Atkins, 1993], but there seems to be no reason against treating them in similar ways as the lexical operations in [Copestake and Briscoe, 1993] or the lexical rules in [Viegas *et al.*, 1996].

Metaphor Extension

The lexical rules that were mentioned before are all based on metonymic relations between lexical meanings. As Apresjan observed already, it is mostly metonymy that leads to systematic (or 'regular') polysemy. Sometimes however, metaphor will also. In view of this, [Martin, 1991] suggests a program of research on sense extension by reasoning with a system of metaphors (MIDAS), which is based on [Lakoff, 1987] and related work.

MIDAS can be used to represent knowledge about conventional metaphors, interpret language using metaphorically based word-senses, and dynamically acquire new word senses as they are encountered during normal processing. Examples of metaphors in MIDAS are (all of them dealing with the UNIX and LISP domains):

(12a) `Container-Metaphor` : How can I get into lisp?

(12b) `Location-Metaphor` : N is at zero.

(12c) `Kill-Metaphor` : How can I kill a process?

2.4.4 Lexical Representations

Instead of stipulating rules that are specific to certain (classes of) lexical entries one could store enough information in the lexical entries themselves and allow a general inference mechanism to generate (context dependent) interpretations from them. This is very much like the *abduction* proposal discussed above apart from the fact that this technique does not assume a specific linguistically motivated representation, where 'linguistic' means: the acknowledgement of lexical semantic patterns among lexical items. Nevertheless, abductive reasoning and lexical semantic representations are in no way mutually exclusive and their combination would be quite fruitful.

One option for specifying lexical (semantic) representations is DATR [Evans and Gazdar, 1996]. A representation in DATR takes the form of a set of *path equations* which are basically attribute-value pairs as in constraint-based approaches [Shieber, 1992]. In this respect lexical entries in DATR are similar to the ones in [Copestake and Briscoe, 1993]. These could in fact be expressed in DATR because it is not restricted to any particular theory. It is a representation language for lexical knowledge with the possibility of defining (multiple) inheritance and defaults.

The representation of systematic polysemy which concerns us here most can be handled in DATR by using *re-entrance* (see e.g.: [Shieber, 1992]). The representation in Figure 2.6 illustrates the systematic polysemy of fruits like cherry, mango, etc. This class of lexical items is systematically polysemous between a **fruit**, the **tree** that grows the fruit and the **wood** produced by the **tree** (from [Evans and Gazdar, 1996], after a proposal in [Kilgarriff, 1995]). By using a re-entrance of the semantic 'glosses' the representation integrates the different aspects of cherry into one comprehensive structure from which each separate interpretation can be generated on request.

```
Cherry:
    <> == NOUN
    <mor root> == cherry
    <sem gloss 1> == sweet red berry with pip
    <sem gloss 2> == tree bearing <sem gloss 1>
    <sem gloss 3> == wood from <sem gloss 2>
```

Figure 2.6: Lexical entry for cherry in DATR, expressing systematic polysemy

2.5 Systematic polysemy of other parts of speech

Central to this thesis is the investigation of systematic polysemy in nouns. However, also other parts of speech show a systematic behavior in respect to their different interpretations. Notably verbs, as in examples (13a) and (13b) ([Pustejovsky and Busa, 1995], see also [Levin, 1993] for an alternative view):

(13a) The enemies sank the boat.

(13b) The boat sank.

Also, adjectives can be systematically polysemous, as in examples (14a) through (14d) ([Pustejovsky and Boguraev, 1993]):

(14a) a fast typist

(14b) a fast game

(14c) a fast book

(14d) a fast driver

Nouns, verbs and adjectives (and adverbs) are all *content* words. However, also non-content or *function* words can be said to be sytematically polysemous. Studies have been done for instance on prepositions [Brugman, 1981]. Even function words with hardly any conceptual content, like connectives have predictable polysemous behaviors. This has been studied mostly in a formal semantic framework, in which the term *polymorphic*, rather than *polysemous* is used [Leiss, 1990] [Morrill, 1992] [Shieber, 1992].

Chapter 3

Generative Lexicon Theory

3.1 Introduction

The previous chapter presented a number of approaches for dealing with systematic polysemy and adjacent phenomena (metonymy, metaphor, sense extension, coercion, etc.) Here I present an integrative theory of lexical semantics that addresses all of these issues by looking not only at lexical meaning by itself, but also at the interface between lexical and compositional meaning; between lexical meaning and syntax; and between lexical meaning and discourse. In this way *generative lexicon* theory ($\mathcal{GL}$) has succeeded in describing and explaining many cases of systematic polysemy. The approach in the thesis is based on this theory.

$\mathcal{GL}$ is both a linguistic theory of semantic interpretation and a theory of lexical semantic knowledge representation. It can be 'decomposed' into several components that are closely related but not fully dependent on each other. The following is an overview of the theory as it emerges from existing literature (see [Pustejovsky, 1991a], [Pustejovsky *et al.*, 1993], [Pustejovsky and Boguraev, 1993], [Pustejovsky, 1994a], [Pustejovsky, 1994b], [Pustejovsky and Busa, 1995], [Pustejovsky, 1995] and [Pustejovsky, 1996]).

Representation

EVENT STRUCTURE	:	Representation of tense and aspect
ARGUMENT STRUCTURE	:	Representation of valency and selection restrictions
QUALIA STRUCTURE	:	Representation of systematic polysemy
QUALIA ROLES	:	Representation language for QUALIA STRUCTURE
DOTTED TYPES	:	Type system for QUALIA ROLES
CO-SPECIFICATION	:	Representation of syntax-semantics mapping

Interpretation

PROJECTIVE INHERITANCE	:	Extended type driven interpretation
CO-COMPOSITIONALITY	:	Bi-directional compositional constraints
COERCION	:	Type error handling

3.2 Representation

3.2.1 Event, Argument and Qualia Structure

In [Pustejovsky, 1991a] and more elaborate in [Pustejovsky, 1995] the semantics of a lexical item is defined by the following tuple:

> $\alpha = \langle \mathcal{A}, \mathcal{E}, \mathcal{Q}, \mathcal{I} \rangle$
>
> '...where $\mathcal{A}$ is the argument structure, $\mathcal{E}$ is the specification of the event structure, $\mathcal{Q}$ provides the binding of these two parameters in the qualia structure, and $\mathcal{I}$ is an embedding transformation, placing α within a type lattice, determining what information is inheritable from the global lexical structure.'

Although all four of the substructures are meant to apply to *all* parts of speech in [Pustejovsky, 1995], qualia structure was originally designed to deal specifically with nouns [Pustejovsky, 1989]. Event and argument structure are longer standing lexical semantic representations that traditionally deal specifically with verbs (and nominalizations). Inheritance structure seems to be more a resulting structure that emerges after applying *projective inheritance* (see Section 3.3.1 below) to the other three structures.

The thesis is only concerned with nouns and I therefore restrict the representation of lexical semantic structure to qualia structure, mainly because from existing literature it is not exactly clear how event and argument structure for nouns would look like. Some aspects of both representations will be included implicitly into the qualia structures I propose and I will therefore discuss them briefly here also.

Event Structure

$\mathcal{GL}$ assumes every lexical item to express some *event*, which is either simple or made up of several sequential or overlapping sub-events. Event structure represents this with 'three primary components: the primitive *event type* of the lexical item; the rules of *event composition*; the *mapping rules* to lexical structure' [Pustejovsky, 1991b]. Events can be classified into three types: STATE (a single event - be sick, love, know), PROCESS (a sequence of events - run, push, drag) and TRANSITION (an event relative to its opposition - give, open, build, destroy). The examples are all verbs, not much is said in the existing literature about event structure for nouns. In Chapter 7 I will discuss some examples of event denoting noun classes and show their representation in qualia structure. However, no separate event structures will be assigned to them.

Argument Structure

A 'minimal specification' of the semantic structure of a predicative lexical item includes a representation of the arguments with which it is associated, or to be more precise: the number of arguments and their mapping onto syntactic structure. The problem of acquiring and representing this knowledge has been addressed by most current linguistic theories. All of them seem to agree on a tight coupling between neccesary arguments on the semantic level and realized syntactic constituents. The nature of this mapping, however, leaves some room for discussion.

In [Pustejovsky, 1995] a further distinction is made between types of arguments, specifically for verbs, that extends argument structure to include not only necessary but also optional arguments. More specifically these include: *default arguments* ('[p]arameters which ... are not necessarily expressed syntactically') and *shadow arguments* ('[p]arameters which ... can be expressed only by operations of subtyping or discourse specification'). I will not elaborate on this any further here, but return to some aspects of the syntax-semantics interface in Chapter 5.

Qualia Structure

[Pustejovsky, 1989] suggests 'that there is a system of relations that characterizes the semantics of nominals, very much like the argument structure of a verb. ... Essentially, the qualia structure of a noun determines its meaning as much as the list of arguments determines a verb's meaning.' In the context of this thesis it displays the systematic polysemy of a lexical item by representing how its different semantic aspects are related to each other. I take *qualia structure* to consist of a set of *lexical* entailments, which is not to be seen strictly as *logical* entailment. The term is used here in the informal sense of [Cruse, 1986]. Entailment in qualia structure could however be given a *default logic* interpretation by adopting the idea of *common sense* entailment [Asher and Morreau, 1990]. For instance, the word book by default entails at least the two events of reading and writing it, besides the knowledge that it consists of several parts: the cover, pages, etc. This should not be confused with the *connotation* of a word. Connotations are rather subjective *associations* among lexical items that are not reflected in linguistic structure, e.g.: the association between my favorite book and the person who gave it to me. Qualia structure on the other hand should be viewed as an *underspecified* representation of the *denotation* of a lexical item[1].

3.2.2 Qualia Roles

In [Pustejovsky, 1991a] four *qualia roles* are introduced, each addressing a different semantic aspect:

FORMAL	:	inheritance (*hyponymy*)
CONSTITUTIVE	:	modification (including *meronymy*)
TELIC	:	purpose
AGENTIVE	:	causality

These four roles are described as 'a necessary component' of semantic representation, but they are nowhere assumed to be exclusive. Some researchers working with $\mathcal{GL}$ have made suggestions for a

[1]For more on *underspecification* see for instance the collection of papers in [van Deemter and Peters, 1996]

revised qualia structure [Calzolari, 1991] [Copestake and Briscoe, 1993], which includes not only these four roles, but any kind of (lexical) knowledge. From such a point of view qualia structure starts to resemble something like *conceptual graphs* [Sowa, 1984] [Sowa, 1993], the *(sub)world concepts* in [Nirenburg and Raskin, 1987] or any other semantic network oriented approach. This may in fact be a fruitful way of using qualia structure and it will be viewed in such a way in the context of this thesis. It is, however, important to stress the mapping between these structures and syntactic structure, which is central in the approach taken in this thesis (see Chapter 5 and Section 3.2.4 below).

3.2.3 Dotted Types

Qualia roles are typed to a specific class of lexical items. Types are either simple (**human, artifact,...**) or complex (e.g., **information•physical**) [Pustejovsky, 1994b]. Complex types are called *dotted types* after the 'dots' that are used as type constructors. Here I introduce two different dots:

'•' connects systematically related types that are *always* interpreted simultaneously

'∘' connects systematically related types that are *never* interpreted simultaneously

Both '$\sigma \bullet \tau$' and '$\sigma \circ \tau$' denote sets of *pairs* of objects $\langle a, b\rangle$, a an object of type σ and b an object of type τ. A condition $a\mathbf{R}b$ restricts this set of pairs to only those for which some relation $\mathbf{R}$ holds, where $\mathbf{R}$ denotes a subset of the Cartesian product of the sets of type σ objects and type τ objects.

$$[\![\sigma \bullet \tau]\!] = \{< a, b > \mid a \in \sigma \textit{ and } b \in \tau \textit{ and } a\mathbf{R}b \textit{ and } < a, b > \textit{ is a complex object } \}$$

$$[\![\sigma \circ \tau]\!] = \{< a, b > \mid a \in \sigma \textit{ and } b \in \tau \textit{ and } a\mathbf{R}b\}$$

Figure 3.1: Definition of '•' and '∘' type constructors

The difference between types '$\sigma \bullet \tau$' and '$\sigma \circ \tau$' is in the nature of the objects they denote. The type '$\sigma \bullet \tau$' denotes sets of pairs of objects where each pair behaves as a *complex* object in discourse structure. For instance, the pairs of objects that are introduced by the type **information•physical** (book, journal, scoreboard, ...) are addressed as the complex objects ⟨**x:information**, **y:physical**⟩ in discourse. On the other hand, the type '$\sigma \circ \tau$' denotes simply a set of pairs of objects that do not occur together in discourse structure. For instance, the pairs of objects that are introduced by the type **form∘artifact** (door, gate, window, ...) are not (normally) addressed simultaneously in discourse, rather one side of the object is picked out in a particular context. Nevertheless, the pair as a whole remains active during processing.

According to this definition, dotted types refer to *tuples* of objects. Two or more semantic sides of a lexical item are implied and may or may not be simultaneously addressed during discourse, depending on the type constructor used ('•' or '∘'). Consider the following examples:

(15a) He threw out the book after reading it.

book: **information•physical**

(15b) Last week I attended his interpretation of Bach's second sonata, the one over there with the picture on the cover.

sonata: **information•physical•event**

(15c) ?? The newspaper that fired its editor fell from the table.

newspaper: **(information•physical)∘organization**

The last example suggests that once the **organization** interpretation has been selected there is no or less access to the **information•physical** interpretation. This could be represented by the type **(information•physical)∘organization**, using both type constructors. Example (16) however shows that such access restrictions can be overruled in puns and other humorous use of language.

(16) 'Freedom of the press is only guaranteed to those who own one.' (A. J. Liebling, 1960)

3.2.4 Co-Specification

An important aspect of the representation of lexical semantic structure in $\mathcal{GL}$ is the mapping onto syntactic structure. In [Pustejovsky, 1991a] and repeated in [Pustejovsky, 1995] an implicit representation of this mapping is proposed by *co-specifying* the verbs that a noun co-occurs with in the TELIC and/or AGENTIVE qualia roles. Consider for instance the representation for cake in Figure 3.2. The AGENTIVE role specifies a verb (to bake) with which this noun co-occurs (by default).

$$\begin{bmatrix} \textsf{cake} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG}1 = \mathbf{x{:}food\_ind} \\ \text{D-ARG}1 = \mathbf{y{:}mass} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \mathbf{y} \\ \text{FORMAL} = \mathbf{x} \\ \text{TELIC} = \mathbf{eat(e_2,z,x)} \\ \text{AGENTIVE} = \mathbf{bake\_act(e_1,w,y)} \end{bmatrix} \end{bmatrix}$$

Figure 3.2: Implicit Co-Specification

In this fashion, co-specification is very much an implicit representation of *collocations*, the structural co-occurrence of two or more lexical items. This could be represented explicitly also, which is the approach taken in DIDEROT [Cowie *et al.*, 1993], an information extraction system that employs lexical semantic representations based on $\mathcal{GL}$. Figure 3.3 is an example lexical entry in this system for the noun agreement. The COSPEC feature gives a number of collocational patterns in which this noun participates. In this case the patterns were acquired from the WALL STREET JOURNAL corpus. The patterns are based on regular expressions, but here `*` implies 'anything' instead of 'repetition', while 'self' stands for the represented lexical item ('agreement' in this case).

```
gls(agreement,
    syn([type(n),
    caseinfo([subcat1(A1), subcat2(A2), subcat3(A3),
              case(A1,np), case(A2,to), case(A3,np)]),
    inflection([pastp(agreed)])]),
    qualia([formal([decision])]),
    args([arg1(A1, syn([type(np)]), qualia([formal([organization])])),
          arg2(A2, syn([type(np)]), qualia([formal([organization])])),
          arg3(A3, syn([type(np)]), qualia([formal([organization])]))]),
    cospec([[formation, of, A2, *, self, *, A1, *, A3],
            [formation, of, A2, *, self, *, A1],
            [A1, *, has,reached,*, self,*, A2, with, A3],
            [A1,and,A3, *, have,reached,*, self,*, A2]]))).
```

Figure 3.3: Explicit Co-Specification

In the context of this thesis I take a combined approach by employing the explicit co-specifications of the DIDEROT approach and mapping them onto implicit co-specifications in the qualia structure by using re-entrance. See Figure 3.4 for an example. Note however, that the syntax-semantics mapping is not straightforward and is highly domain dependent. Chapter 5 describes in more detail how corpus-based co-specification patterns can be mapped onto qualia structure.

$$
\begin{bmatrix}
\textsf{agreement} & \\
\textsc{qualia} = & \begin{bmatrix}
\textsc{formal} = \mathbf{x{:}event} \\
\textsc{constitutive} = \mathbf{y{:}event} \\
\textsc{telic} = \\
\quad \mathbf{form(e_1,x,a_2{:}organization,(a_1{:}organization \wedge a_3{:}organization))} \\
\quad \mathbf{reach(e_1,a_2{:}organization,(a_1{:}organization \wedge a_3{:}organization),x)}
\end{bmatrix} \\
\textsc{cospec} = & \\
& \texttt{[formation, of, } a_2\texttt{, *, x, *, } a_1\texttt{, *, } a_3\texttt{]} \\
& \texttt{[formation, of, } a_2\texttt{, *, x, *, } a_1\texttt{]} \\
& \texttt{[}a_2\texttt{, *, has,reached,*, x,*, } a_1\texttt{, with, } a_3\texttt{]} \\
& \texttt{[}a_1\texttt{,and,}a_3\texttt{, *, have,reached,*, x,*, } a_2\texttt{]}
\end{bmatrix}
$$

Figure 3.4: Combined Co-Specification

3.3 Interpretation

The interpretation of signs, such as lexical items in natural language, is commonly understood as linking the name of the sign to an object in the world, or to a representation of that object, depending on one's metaphysical views. Considering the fact that NLP systems are still restricted to having access only to *representations* of objects in the world I will stick with the latter approach. To be more precise, in the context of this thesis interpretation means linking lexical items to representations of (classes of) objects. Or to be even more precise, interpretation here means assigning lexical semantic structure to lexical items. This involves inferential reasoning and handling the constraints on that. This section discusses the $\mathcal{GL}$ approach to infer knowledge structures by way of inheritance and the constraints on those inferences that arise from compositional semantics and discourse.

3.3.1 Projective Inheritance

Organization of knowledge in hierarchies allows for set based inferences as discussed in [Touretzky, 1986]. Many of the systems that have been developed within the AI community have a basic architecture that is based on inheritance as the primary inference engine between knowledge structures in the system. Inheritance in such systems is typically represented through `isa` links that indicate *set inclusion* and `hasa` links that represent *attributes* of the object under description.

Qualia structure can be viewed as a collection of such `isa` and `hasa` links where the FORMAL role is taken as an `isa` link that represents *set inclusion*, while the other three roles represent distinguishing attributes. With traditional inheritance only the `isa` link carries inferential knowledge *directly*. Attributes can only be inherited among objects through their `isa` links, that is through their inclusion into sets.

In [Pustejovsky and Boguraev, 1993] an alternative view of inheritance is proposed that would allow inferences between knowledge structures not only through the `isa` links but also through the `hasa` links that are actually to be upgraded into a kind of 'role specific' `isa` links. In fact the four qualia roles are to be interpreted as introducing four separate inheritance lattices: `is_formal, is_constitutive, is_telic, is_agentive`. Inheritance can now proceed through four different 'channels' that are mutually independent, which [Pustejovsky and Boguraev, 1993] calls *projective* inheritance. Figure 3.5 gives an example that explains the difference in meaning between dictionary and play, both inheriting (partly) from book.

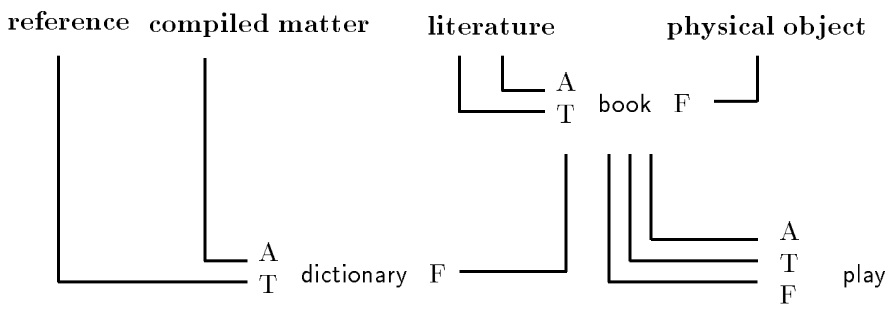

Figure 3.5: Projective Inheritance for dictionary and play

The example shows that a dictionary and a play are both a kind of book, but in different ways. A dictionary is only a book in respect to it being a **physical object** which it inherits through the FORMAL role. Its TELIC and AGENTIVE roles differ from that of book and are not inherited. In contrast, a play inherits all aspects of book, through the FORMAL, TELIC and AGENTIVE roles. This would explain why one can 'read' a book or a play, but not a dictionary. Also, a dictionary is not 'written' in the same way a book or a play are, but is rather 'compiled' out of previously written material.

Projective Inheritance and Dotted Types

The introduction of *projective inheritance* in [Pustejovsky and Boguraev, 1993] historically preceeds that of *dotted types* in [Pustejovsky, 1994b] and is partly superseded by it. With dotted types, the FORMAL role can be made more complex by itself and allows for more complex inferences through (traditional) inheritance. Nevertheless there remains a need for explaining why dictionaries and plays are both books but not in the same way, although both of them inherit the **information•artifact** type from book. The analysis in Figure 3.6 shows how dotted types *and* projective inheritance can be used to explain this:

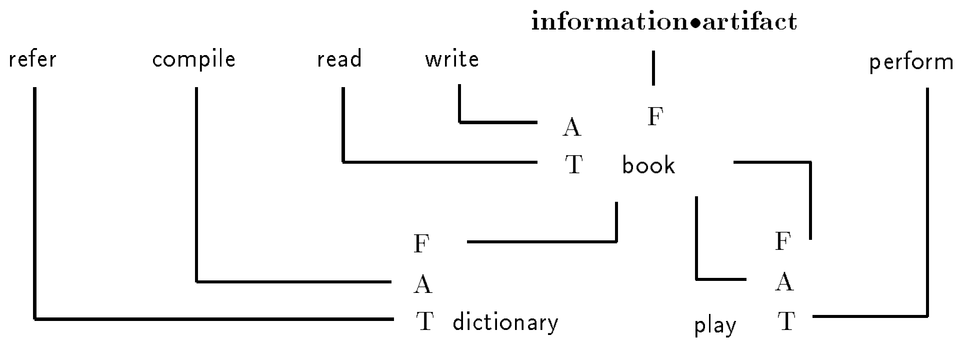

Figure 3.6: Projective Inheritance and Dotted Types for dictionary and play

Unified Types

Figure 3.6 actually implies two kinds of complex types as they are discussed in [Pustejovsky, 1995]. Next to the dotted types one can also distinguish *unified types* that unify qualia structures of different but not conflicting types. That is, one can unify qualia structures of different types only if their FORMAL roles have non-conflicting types, themselves simple or dotted.

Figure 3.7 illustrates this with an even more fine grained analysis of the same example. Although the noun 'play' is into some respect an **information•artifact** like a book is, it is also an **event** that can be 'performed', which makes it of type **(information•artifact)•event**. The AGENTIVE role for play (to write) is inherited from book, whereas the TELIC role (to perform) is inherited directly from the **(information•artifact)•event** type in the FORMAL role. The noun play can thus be said to be of the unified type: $\tau_1 \sqcup \tau_2$ (**information•artifact**$\sqcup$**(information•artifact)•event**)

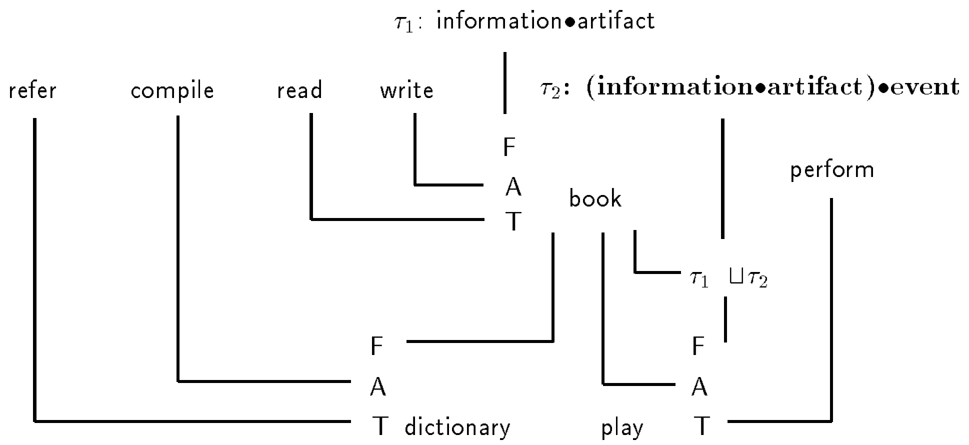

Figure 3.7: Unified and Dotted Types for dictionary and play

The approach in this thesis employs only dotted types. The overlap in semantics between the type **information•artifact** and **(information•artifact)•event** (to write a play) that can be expressed by *unifying* those types is already implicit in their dotted structure. In fact, unified types are subsumed by dotted types as a special case, more specifically by dotted types constructed with the '•' operator. The '∘' operator connects types that are not unifiable by definition.

The type $\sigma \bullet \tau$ denotes a set of *unified* objects if σ and τ are unifiable, or a set of pairs of objects otherwise. Unified objects do not constitute pairs of objects, but are complex objects constructed out of objects of their unifying types. An extended definition of both the '•' and '∘' type constructors along these lines is given in Figure 3.8.

$$[\![\sigma \bullet \tau]\!] = \{\ a \sqcup b \mid a \in \sigma \ and \ \ b \in \tau \ and \ a\mathbf{R}b \ and \ a \sqcup b \ \neq \perp\}$$

$$[\![\sigma \bullet \tau]\!] = \{< a, b > \mid a \in \sigma \ and \ \ b \in \tau \ and \ a\mathbf{R}b \ and \ < a, b > \ is \ a \ complex \ object \ \}$$

$$[\![\sigma \circ \tau]\!] = \{< a, b > \mid a \in \sigma \ and \ \ b \in \tau \ and \ a\mathbf{R}b\}$$

Figure 3.8: Extended definition of '•' and '∘' type constructors

An Example: artifact•tool

In [Pustejovsky, 1995] an example of unified types is given that can be reconstructed in terms of the definition just given. Consider the following qualia structures for the types **artifact** and **tool**:

$$\begin{bmatrix} \textsc{formal} = \mathbf{y{:}artifact} \\ \textsc{agentive} = \mathbf{make(e,x,y)} \end{bmatrix} \qquad \begin{bmatrix} \textsc{formal} = \mathbf{y{:}tool} \\ \textsc{telic} = \mathbf{Q(e,x,y)} \end{bmatrix}$$

Figure 3.9: Qualia structures for **artifact** and **tool**

The **artifact** type is unspecified for the TELIC role while the **tool** type is unspecified for the AGENTIVE role, which makes them unifiable into the dotted type **artifact•tool** as given in Figure 3.10. This type includes tools that are created by man (artifacts) like 'knife', but excludes tools that are found in nature (natural objects).

$$\begin{bmatrix} \textsc{formal} = \mathbf{y{:}artifact{\bullet}tool} \\ \textsc{agentive} = \mathbf{make(e,x,y)} \\ \textsc{telic} = \mathbf{Q(e,x,y)} \end{bmatrix}$$

Figure 3.10: Qualia structure for **artifact•tool**

The example shows that unified types are really a special case of dotted types and do not employ projective inheritance. On the other hand, non-straightforward inferences as enabled by projective inheritance seem useful in their own right as shown in the next section.

3.3.2 Coercion, Polymorphism and Co-Compositionality

Coercion and Polymorphism

$\mathcal{GL}$ assumes some form of type driven interpretation, which by now is a standard approach in compositional semantics, following Montague's influential work [Montague, 1974]. Compositionality in this model is typically viewed as a *functor* taking any number of *typed* arguments. Application of the functor to the argument(s) can only be successful if the arguments are of the 'right' type as requested by the functor. More flexible approaches allow arguments to 'shift' their type according to some rule or general principle. This includes the use of *polymorphic types* [Partee, 1987] [Moortgat, 1988] [Leiss, 1990] [Shieber, 1992] [Morrill, 1992] and *coercion* [Moens and Steedman, 1987] [Pustejovsky, 1991a] [Hobbs *et al.*, 1993] [Pollard and Sag, 1994].

Polymorphism and coercion are really two sides of the same coin. An argument that is typed to be of some polymorphic type can assume different 'forms' (hence *poly-morphism*) depending on which particular functor has been applied to it. Coercion is taking things from the perspective of the functor, that 'coerces' its argument(s) to take on a particular 'form' [Cardelli and Wegner, 1985].

Most of the applications of type shifting as proposed in the literature are semantically motivated approaches to syntactic problems. Consider for instance the following sentence:

(17) John and every (other) student went to her party.

In Montague's type system ($_e$ for individual; $_t$ for proposition) [every (other) student]$_{NP}$ is of type $_{<<e,t>,t>}$, whereas [John]$_{PN}$ is of type $_e$. Although of different type they are coordinated on the same level by the conjunction operator (functor) 'and'. One can solve this problem by shifting [John]$_{PN}$ from type e to type $_{<<e,t>,t>}$ [Partee, 1987]. The semantical motivation behind this is that 'John' can be interpreted as the set of properties of 'John', which is exactly expressed by the type $_{<<e,t>,t>}$: a function from sets of properties ($_{<e,t>}$) to truth values ($_t$).

Co-Compositionality

In $\mathcal{GL}$ coercion and polymorphic types are combined with a much richer lexical semantic structure than normally assumed in Montague Grammar or similar approaches, which allows for deeper analyses of type shifting solutions to syntactic problems and opens up possibilities for explaining additional problems that are more semantic in nature[2]. Consider these examples:

(18a) Mary wants [John to leave]$_S$.

(18b) Mary wants [to leave]$_{VP}$.

(18c) Mary wants [a beer]$_{NP}$.

Assuming the semantics of the verb 'want' does not change, the different types of argument in object position (S, VP and NP) have to be shifted to a common underlying type. In fact, all three arguments are to be interpreted as propositions of type t, which can be analysed as in (19a) through (19c), assuming the following type assignments:

Mary, John	:	e	Proper Name
(to) leave	:	$<e,t>$	Intransitive Verb
(to) drink	:	$<e,<e,t>>$	Transitive Verb
(to) want	:	$<t,<e,t>>$	Aspectual Transitive Verb
a beer	:	$<<e,t>,t>$	NP

(19a) Mary$_e$ wants$_{<t,<e,t>>}$ [John to leave]$_t$.

(19b) Mary$_e$ wants$_{<t,<e,t>>}$ [*Mary*$_e$ [to leave]$_{<e,t>}$]$_t$.

(19c) Mary$_e$ wants$_{<t,<e,t>>}$ [*[Mary*$_e$ *[to drink]*$_{<e,<e,t>>}$*]*$_{<e,t>}$ [a beer]$_{<<e,t>,t>}$]$_t$.

$\mathcal{GL}$ explains the type shifts involved by employing a richer lexical semantic structure. The shift of [to leave]$_{VP}$ from $<e,t>$ to t can be explained simply by taking into account the implicit 'agent' of this intransitive verb. The shift of [a beer]$_{NP}$ is somewhat harder to explain, but involves a deeper analysis of the lexical semantic structure of the noun 'beer'. To be more precise, its qualia structure should represent a number of events in which this noun typically participates (e.g. 'to drink').

This example shows that the analysis of such problems not only employ the lexical semantic structure of the functor (to want), but also that of the (*head of the*) argument (a beer). As such, the analysis goes

[2]For instance the *bridging* problem in discourse analysis [Bos *et al.*, 1995].

beyond traditional composition in semantic analysis and is therefore referred to as *co-compositonality* (see [Pustejovsky, 1991a], [Buitelaar and Mineur, 1994], [Pustejovsky, 1995], [Mineur and Buitelaar, 1996]). The type shifts that are involved in this are called *coercions*, which is rather an extension of the technical term by that same name, because they are more like transfers of type than 'traditional' type shifts that can only be employed according to certain type-theoretic rules.

Coercion as defined in [Pustejovsky, 1995], but also in [Hobbs *et al.*, 1993] – although with rather different theoretical assumptions – is more concerned with the knowledge that is involved in the analysis of such problems than with the formal limits of its application. However, given a rich enough type theory that includes sortal restrictions on types, reflecting conceptual categories like **artifact, human,...**, one could perhaps give a formal definition of coercion along the lines of Montagovian type theory that covers examples brought up already by [Nunberg, 1979] and extended by [Pustejovsky, 1991a], [Pustejovsky, 1995] and [Hobbs *et al.*, 1993]. Experiments along these lines have been reported in [Dölling, 1994] and [Verschuur, 1993]. This thesis by itself is not concerned primarily with issues of formal interpretation and considers coercion merely as a 'selection' operation that picks out appropriate knowledge as stored in the qualia structure on request.

Coercion and Qualia Structure

To illustrate the influence of qualia structure on the analysis of sentences like (19c) consider the following, similar examples:

(20a) He began [a new book]$_{<<e,t>,t>}$.

(20b) He began [a new dictionary]$_{<<e,t>,t>}$.

The verb 'to begin' is aspectual and takes an argument of type $_{t}$. In both examples however it has arguments of type $_{<<e,t>,t>}$, that need to be coerced accordingly. As in example (19c), the NP's are re-interpreted as 'ellipsed' propositions. The qualia structure now supplies the specific verb(s) to assume in each case, for 'book' this will be something like 'to read' or 'to write', whereas 'dictionary' is more likely to introduce 'to consult' or 'to compile'. This kind of knowledge is however highly domain dependent – taking the form of corpus specific *collocations* (see Section 3.2.4) – and is further constrained by discourse structure [Lascarides and Copestake, 1995] [Verspoor, 1996].

The use of qualia structure in semantic interpretation is illustrated maybe even clearer in examples that involve systematic polysemy as in sentence (43a):

(21) BMW announced a new model.

Company names like 'BMW' are systematically polysemous between the organizations they represent, the buildings they occupy and the people they employ. In this case it is the latter interpretation: people within the company normally do announcements. As such, the 'group of **human** beings' interpretation of 'BMW' must be selected out of the qualia structure, which corresponds with the selection restriction of the verb 'to announce' that takes an NP of sort **human** in subject position.

The thesis is concerned mostly with examples of this kind and is aimed at providing natural language processing systems with the appropriate knowledge for handling them as will be shown in the remaining chapters.

Chapter 4

CoreLex: An Ontology of Systematic Polysemous Classes

4.1 Introduction

In this chapter I describe the structure and content of an ontology (CoreLex) that builds on the assumptions about lexical semantics outlined above. More specifically, CoreLex is to be 'structured in such a way that it reflects the lexical semantics of a language in systematic and predictable ways' [Pustejovsky *et al.*, 1995]. This assumption is fundamentally different from the design philosophies behind existing lexical semantic resources like WordNet that do not account for any regularities between senses[1]. The strategies for deriving such an ontology of systematic polysemous classes from WordNet will be explained throughout this chapter, but can be summarized by the following three stages: 1. Reducing WordNet senses to a set of 'basic types'; 2. Organizing the basic types into systematic polysemous classes, that is, grouping together lexical items that share the same distribution of basic types; 3. Representing systematic polysemous classes through *underspecified semantic type* definitions, that extend into *qualia structure* representations. Each one of these steps will be explained in detail below, but first I will take a closer look at WordNet.

[1] WordNet 1.5 in fact includes an experiment in representing related senses, called *cousins*. The scope of this experiment is however small and its results are very preliminary

4.2 WordNet

The most widely used lexical database in NLP research does not have an origin in linguistics, but rather in psychology, or psycholinguistics to be precise. WordNet seems primarily to be designed as a computational account of the human capacity of linguistic categorization [Miller *et al.*, 1990]. It therefore covers a rather extensive set of lexico-semantic categories, or *synsets* as they are commonly referred to[2]. Synsets are collections of synonyms, that are used to group together lexical items according to similar meanings. For instance, a board and a plank are similar lexical items and can thus be grouped together in the synset: {**board, plank**}. At the same time such synsets can be used to disambiguate between homonymic senses. The noun board not only refers to a piece of lumber, but can be used also to designate a group of people. Consequently, this noun is grouped in two synsets, once in its 'lumber' meaning ({**board, plank**}) and once in its 'people' meaning ({**board, committee**}). So, in fact synsets are not really made up of *lexical items*, but rather of *lexical meanings* or *senses*. Observe, that the synset strategy defines lexical meaning in an *implicit* way, contrary to the explicit definitions that are found in traditional dictionaries.

Synsets range from the very specific to the very general. Very specific synsets typically cover only a small number of lexical items, while very general ones tend to cover many. For instance, the synset {**exemption, immunity, granting_immunity**} only covers: exemption, immunity and fix, while the synset {**object, inanimate_object, physical_object**} covers thousands of lexical items. Moreover, a hierarchical ordering between synsets exists, with general synsets covering more specific ones. For instance, the synset {**life_form, organism, being, living_thing**} covers the more specific ones: {**person, individual, someone, mortal, human, soul**}, {**animal, animate_being, beast, brute, creature, fauna**} and {**plant, flora, plant_life**}[3]. On its turn, {**animal, animate_being, beast, brute, creature, fauna**} covers the more specific synsets: {**chordate**}, {**invertebrate**} and {**acrodont**}. The {**chordate**} synset covers {**vertebrate, craniate**}, {**cephalochordate**}, {**tunicate, urochordate, urochord**} and so on.

[2] There are 91591 synsets in total for the WordNet 1.5 release, of which 60557 for nouns.

[3] Plus some smaller ones like {**microorganism**}

The organization of WordNet through lexical meaning, rather than through the lexical items themselves makes it different from traditional dictionaries and more like a thesaurus. An additional difference is the division into four separate databases, one for each major part of speech: *verbs, nouns, adjectives* and *adverbs*. The motivation for this design decision stems from psycholinguistic research on word association. Apparently, human subjects tend to associate among words mainly within the same syntactic category. '[N]oun probes elicited nouns responses 79% of the time, adjectives elicited adjectives 65% of the time, and verbs elicited verbs 43% of the time' [Miller *et al.*, 1990] (quoting from [Fillenbaum and Jones, 1965]). The inherent redundancy of this approach, with lexical items showing up in more than one database, is taken for granted because of its psychological realism.

Each database is organized in a different way. Nouns are organized as topical hierarchies, verbs by a variety of entailment relations and adjectives and adverbs as N-dimensional hyperspaces. For the nouns database a set of 'unique beginners' or *top types* is distinguished. These are synsets that are not covered by any other synset. Each one of them constitutes a separate hierarchy corresponding to 'relatively distinct semantic fields, each with its own vocabulary'. In addition, such a partitioning 'also has practical advantages: it reduces the size of the files that the lexicographers must work with, and makes it possible to assign the writing and editing of different files to different lexicographers.' According to [Miller *et al.*, 1990] this set consists of 25 types that are listed in Figure 4.1.

{ **act, action, activity** }	{ **natural object** }
{ **animal,fauna** }	{ **natural phenomenon** }
{ **artifact** }	{ **person, human being** }
{ **attribute, property** }	{ **plant, flora** }
{ **body, corpus** }	{ **possession** }
{ **cognition, knowledge** }	{ **process** }
{ **communication** }	{ **quantity, amount** }
{ **event, happening** }	{ **relation** }
{ **feeling, emotion** }	{ **shape** }
{ **food** }	{ **state, condition** }
{ **group, collection** }	{ **substance** }
{ **location, place** }	{ **time** }
{ **motive** }	

Figure 4.1: Top types in WordNet

In fact, there are only 11 of such types identifiable in WordNet [4], although the set of 25 seems to be used in compiling the database. Each one of these 11 *top types* constitutes a separate *ontological domain*, that can be further subdivided as shown in Figure 4.2 below.

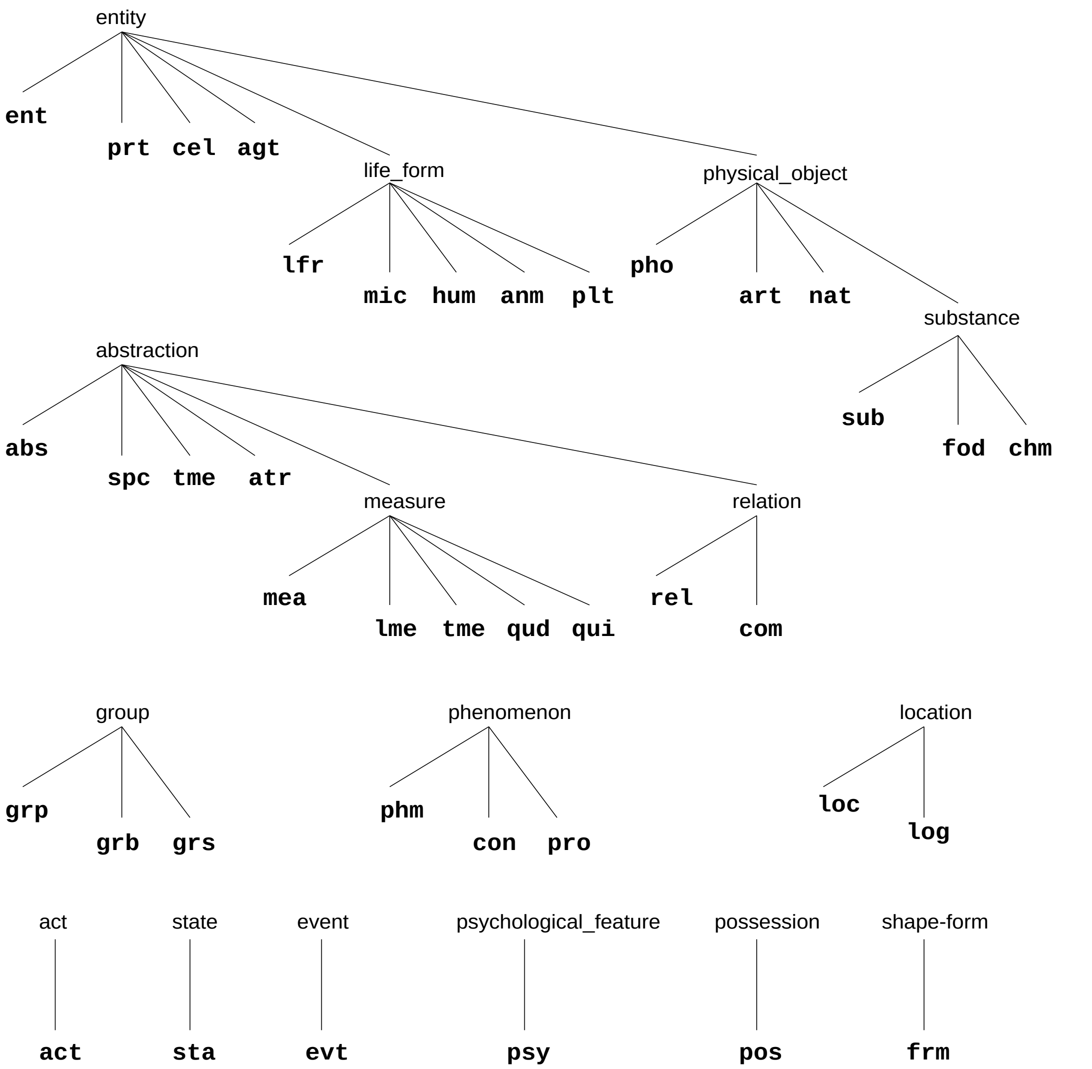

Figure 4.2: The hierarchies of basic types in WordNet

The hierarchy displayed here is used in deriving CoreLex. It extends the 11 WordNet top types with 28 further ones on several sublevels. Together they constitute a set of 39 *basic types*[5]. Figure 4.3 lists each of them together with their frequency in the nouns database, that is, how many noun

[4] This is confirmed by the WordNet 1.5 *'man page'*.

[5] The basic type `'tme'` (**time**) occurs twice, grouping together two subdomains of WordNet as one in CoreLex.

instances there are for each type. Basic types marked by '$*$' are *residual*, meaning their frequencies include only those nouns that are strictly only of that (super)type and do not belong to any of its subtypes. For instance, ABS$*$ has 8 instances that are defined by WordNet as belonging only to the supertype **abstraction** and which are not further specified into a subtype like **definite_quantity** or **linear_measure**.

Type	Corresponding Synsets	Freq
ART	artifact, artefact	8283
ACT	act, human_action, human_activity	6606
HUM	person, individual, someone, mortal, human, soul	6303
GRB	biological_group	4933
ATR	attribute	4137
PSY	psychological_feature	3456
COM	communication	3336
ANM	animal, animate_being, beast, brute, creature, fauna	2703
PLT	plant, flora, plant_life	2311
STA	state	2266
FOD	food, nutrient	1541
LOG	region 1 (geographical location)	1282
NAT	natural_object - water,.. - land,...	1277
SUB*	substance, matter	1189
EVT	event	1082
PRT	part, piece	992
GRS	social_group - people	940
QUD	definite_quantity	777
PRO	process	773
CHM	compound, chemical_compound - chemical_element, element	699
TME	time_period, period, period_of_time, amount_of_time - time_unit, unit_of_time - time	628
AGT	causal_agent, cause, causal_agency	624
POS	possession	571
LOC*	location, (any other location)	567
REL*	relation	506
FRM	shape, form	420
GRP*	group, grouping (any other group)	345
PHM*	phenomenon	342
QUI	indefinite_quantity	295
PHO*	object, inanimate_object, physical_object	186
MIC	microorganism	178
LME	linear_measure, long_measure	100
LFR*	life_form, organism, being, living_thing	61
CEL	cell	57
MEA*	measure, quantity, amount, quantum	38
ENT*	entity	28
CON	consequence, effect, outcome, result, upshot	21
SPC	space	21
ABS*	abstraction	8

Figure 4.3: Basic types in WordNet and their frequencies

4.3 CoreLex

Although WordNet is a valuable resource of lexical semantic information, which is widely used for all kinds of NLP research projects and applications, it has the major omission of a principled distinction between *polysemy* and *homonymy*. All senses in WordNet are at the same theoretical level, without any special attention for senses that might be (systematically) related. For instance, WordNet assigns to the noun book the following senses:

book	1. {publication}	⇒ **artifact**
	2. {product, production}	⇒ **artifact**
	3. {fact}	⇒ **communication**
	4. {dramatic_composition, dramatic_work}	⇒ **communication**
	5. {record}	⇒ **communication**
	6. {section, subdivision}	⇒ **communication**
	7. {journal}	⇒ **artifact**

Figure 4.4: WordNet senses for the noun book

At the top of the WordNet hierarchy these seven senses can be reduced to two basic types: the content that is being communicated (**communication**) and the medium of communication (**artifact**). More accurately, book should be assigned a representation which implies both of these interpretations and connects them to each of the more specific senses that WordNet assigns: **facts, drama** and a **journal** can be *part-of* the content of a book; a **section** is *part-of* both the content and the medium; **publication, production** and **recording** are all *events* in which both the content and the medium aspects of a book can be involved.

Such an approach would imply a two-stage reduction of the WordNet sense assignment. The first stage reduces the bottom level senses to a set of basic types, while the second stage connects these basic types into an integrated type definition, which will in essence be *underspecified* for a specific component 'basic' type. The reduction of the first stage is not strictly necessary. One could also connect the bottom level senses into an underspecified representation. However, given the obviously very large number of bottom level senses this would be highly unpractical.

An important advantage of the CoreLex approach is also more consistency among the assignments of lexical semantic structure. Consider the nouns door, gate and window. Obviously these are similar lexical items, all three of them describing an object that is at the same time a 'movable_barrier' and an 'opening', both in a physical and conceptual way. In WordNet these related meanings are not assigned consistently, as shown in Figure 4.5.

door	1.	{movable_barrier}	⇒ **artifact**
	2.	{entrance, entranceway, entryway}	⇒ **artifact**
	3.	{entree, access, admittance}	⇒ **psychological_feature**
	4.	{house}	⇒ **artifact**
	5.	{room}	⇒ **artifact**
gate	1.	{movable_barrier}	⇒ **artifact**
	2.	{computer_circuit}	⇒ **artifact**
	3.	{gross_income, gross, revenue, receipts}	⇒ **possession**
window	1.	{opening}	⇒ **artifact**
	2.	{panel}	⇒ **artifact**
	3.	{display}	⇒ **artifact**

Figure 4.5: WordNet senses for the nouns door, window and gate

- 'door' is assigned **artifact** and **psychological_feature**, with the 'opening' sense being treated as an **artifact** reading[6]

- 'gate' is assigned **artifact** and **possession** although sense 2. derives from a conceptual interpretation of gate as an opening, and sense 3. from gate as a physical opening

- 'window' is assigned **artifact**, while only sense 2. can be regarded as such, with sense 1. the physical opening and sense 3. the conceptual interpretation of a window

In the CoreLex approach, these nouns are given the same semantic type, which is *underspecified* for any specific 'sense' but assigns them consistently with the same *basic* lexical semantic structure that expresses the regularities between all of their interpretations.

[6]Note the somewhat arbitrary assignment of senses 4. (house) and 5. (room). These derive from such usage as: 'The woman next door'

4.3.1 Systematic Polysemous Classes

We can arrive at classes of systematically polysemous lexical items by investigating which items share the same senses and are thus polysemous in the same way. This section describes how systematic polysemous classes can be derived from WordNet[7] Here I only consider classes of *nouns*, but the process described can be applied to the other parts of speech as well.

We start by reducing the WordNet sense assignments to the set of 39 basic types in CoreLex. For instance, the seven different senses that WordNet assigns to book can be reduced to the two basic types: `'art com'`. We do this for each noun and then group them into classes according to their combination of basic types. This gives a total of 1648 classes. From these we can filter out those classes that have only one member because they obviously do not represent a systematically polysemous *class*.

Such lexical items have a highly idiosyncratic behavior and are most likely homonyms. An example is the noun amazon, the WordNet sense assignment of which is given in Figure 4.6. The four WordNet senses are reduced to the basic types: `'anm hum nat psy'`. Although one could argue that most or all of these are somehow related, this specific combination of basic types is unique to the noun amazon (in WordNet) and is therefore disregarded.

amazon	1.	{woman, adult_female}	⇒ **human**
	2.	{mythical_being}	⇒ **psychological_feature**
	3.	{river}	⇒ **natural_object**
	4.	{bird}	⇒ **animal**

Figure 4.6: WordNet senses for the noun amazon

The most polysemous class is: `'anm art com evt frm hum lme loc nat phm pro prt psy'`. This combination of 13 different basic types is unique to the noun head, which is assigned 23 senses by WordNet as given in Figure 4.7. A number of these are clearly related, indicating that this noun is in fact much less 'ambiguous' than represented here. Nevertheless, because of its idiosyncratic

[7] I also considered Roget as a possible resource, but WordNet is better, because it is more extensive and has a typing system which is less baroque.

behavior it is disregarded also in developing a set of systematic polysemous classes[8].

head	1.	{body_part}	⇒ **part**
	2.	{leader}	⇒ **human**
	3.	{mind}	⇒ **psychological_feature**
	4.	{top}	⇒ **location**
	5.	{point}	⇒ **form**
	6.	{linear_unit}	⇒ **linear_measure**
	7.	{plant_organ}	⇒ **natural_object**
	8.	{pressure}	⇒ **phenomenon**
	9.	{principal}	⇒ **human**
	10.	{drug_user}	⇒ **human**
	11.	{fountainhead}	⇒ **location**
	12.	{pass}	⇒ **event**
	13.	{headway}	⇒ **event**
	14.	{mark}	⇒ **communication**
	15.	{subject}	⇒ **communication**
	16.	{line}	⇒ **communication**
	17.	{striker}	⇒ **artifact**
	18.	{toilet}	⇒ **artifact**
	19.	{projection}	⇒ **artifact**
	20.	{obverse}	⇒ **artifact**
	21.	{membrane}	⇒ **artifact**
	22.	{domestic_animal}	⇒ **animal**
	23.	{oral_sex}	⇒ **process**

Figure 4.7: WordNet senses for the noun head

Removing classes with only one member leaves a set of 529 polysemous classes, of which Figure 4.8 gives some examples. Among these are a number of classes of systematically polysemous nouns that are to be expected given the literature that was discussed in Chapter 2 above. For instance, the classes `'nat plt'` and `'fod nat'` are related to the *grinding* rule. Other classes are less expected, but seem quite intuitive. The class `'art atr sub'` for instance includes a number of nouns (chalk charcoal daub fiber fibre tincture) that refer to an object that is at the same time an **artifact** made of some **substance** that is also an **attribute**. Still other classes simply arise from the WordNet sense assignments and the choice of basic types. An example of this is the class `'lme qud'`, which includes a number of nouns that are assigned both the sense of a **linear_measure** (`'lme'`) and a **definite_quantity** (`'qud'`).

[8] Although we can still relate parts of this polysemous class to a number of *underspecified semantic types*, as explained further in Section 4.4 below.

act anm art	bat drill fly hobby ruff solitaire spat
act art frm grs prt	ball fold
act art grs	artillery assemblage assembly band battery church concourse dance gathering institution market masonry masquerade promenade transportation turnout
act art hum	anachronism chair jumper whip
act art log	berth habitation mooring
act atr evt psy	appearance concentration decision deviation difference impulse outrage variance
act evt rel	blend competition contraction transformation
act log	bolivia caliphate charleston chicago clearing emirate michigan prefecture repair santiago wheeling
act plt	chess grapevine rape
anm nat	alligator broadtail chamois cimarron cob ermine lapin leopard marabou muskrat otter raccoon roan sierra squirrel
art atr sub	chalk charcoal daub fiber fibre tincture
atr plt	carnation fuchsia heather honesty maize pink thrift violet
atr pos	asset birthright bounty cost equity gain largess mortmain profit spirituality toll
com nat	algol baltic creek crux delta lemma sheepskin wig
com psy sta	condition confidence fact mimesis peace problem reason title
evt grp	clump conglomeration overspill
fod nat	berry carrion cocoyam garbanzo java madeira milk mocha sprout
frm prt	convolution crotch excrescence fissure orb pore proboscis sinusoid spicule spine stoma
grp log	garden grove patriarchate scrapheap shrubbery
hum qud	gal gauss gilbert mao newton ohm para quint quintuplet welterweight
lme qud	em fathom fthm inch mil
loc psy	bourn bourne demarcation fairyland rubicon trend vertex viewpoint
nat plt	algarroba almond andromeda anise baneberry butternut candlenut cardamom cardamon cashew chinquapin chokecherry chrysanthemum cobnut cowage cumin filbert ginseng groundnut hazelnut horsechestnut magnolia mandrake marasca orris osier peanut pignut pistachio senega shagbark shellbark squill sugarcane
pos qud	cent centime dividend gross penny real shilling
pos sta	debt dominion indebtedness indemnity subsistence terms
prt qui	breast thigh
psy rel	basis chemistry correlation derivative determinative modality mph paradigm particular quotient
qud tme	century decade half last quaternary teens
rel sta	baronetcy connectedness context efficiency relationship viscountcy
sta tme	adolescence adulthood babyhood bachelorhood childhood day drought eternity gestation infancy latency lull maturity moratorium phase reign sleep widowhood

Figure 4.8: A selection of polysemous classes

Not all of the 529 classes are *systematically* polysemous. Some are collections of homonyms that are *ambiguous* in similar ways, but do not lead to any kind of predictable polysemous behavior, for instance the class `'act anm art'` with the nouns: bat drill fly hobby ruff solitaire spat. Other classes consist of both homonyms and systematically polysemous nouns like the class `act log`, which includes caliphate, clearing, emirate, prefecture, repair, wheeling vs. bolivia, charleston, chicago, michigan. Whereas the first group of nouns expresses two separated but related meanings (the **act** of clearing, repair, etc. takes place at a certain **location**), the second group expresses two meanings that are not related (the charleston dance which was named after the town by the same name). These mixed classes of *ambiguous* and *polysemous* lexical items are to be weeded out carefully, while the *homonymic* classes need to be removed altogether.

4.3.2 Underspecified Semantic Types

The next step in the research is to give type definitions that describe the systematic polysemous classes in an intrinsic way. For instance, in a trivial case, the (basic) type **act** describes those nouns that fall into the class `'act'`. A more complex example includes a number of classes, for instance the type **anp** describes nouns that fall in both classes `'anm'` and `'psy'`, that is, nouns describing *'animals of a psychological, or conceptual nature'* (mythological creatures). These type definitions can be said to be underspecified for a particular class and are therefore referred to as *underspecified semantic types*.

In theory, each underspecified semantic type should correspond to one systematic polysemous class (`'anm psy'` in this case). This would mean that the WordNet analysis as described above would give us 529 types. However, the WordNet sense assignments cannot always be trusted as shown earlier and therefore we often have to take a number of systematic polysemous classes together into a single type definition, because the differences between those classes are not identifiable. An example of this is the type `acr`, corresponding to four different classes that are given in Figure 4.9.

act atr rel	acceleration comparison
act evt rel	blend competition contraction transformation
act rel	abbreviation dealings designation discourse gait glide likening negation neologism neology prevention qualifying reciprocity sharing synchronisation synchronization synchronizing
act rel sta	coordination gradation inclusion involvement

Figure 4.9: Polysemous classes and instances for the type: `acr`

The underspecified semantic type '`acr`' binds together five basic types: **act**, **attribute**, **event**, **relation**, **state**. We can further specify the relations between these through the use of the two dotted type constructors that were introduced in Chapter 3. According to the definitions given there, the type '`acr`' can be represented as **act•attribute•event•relation•state**, introducing a set of tuples of objects that are likely to be addressed in discourse as one complex object. For example, in sentences (22a) and (22b), that were taken from the BROWN corpus, 'delicate transformation' and 'proper coordination' address simultaneously the **event** and the **act** of transforming/coordinating an object R_1, the transforming/coordinating **relation** between two objects R_2 and R_3 and the **attribute** and **state** of this transformation/coordination itself.

(22a) Soon they all were removed to Central Laboratory School where their delicate transformation began.

(22b) Revise and complete wildlife habitat management and improvement plans for all administrative units, assuring proper coordination between wildlife habitat management and other resources.

Underspecified semantic types in the *abstraction* domain

The type '`acr`' spans three ontological domains in the WORDNET hierarchy: **abstraction**, **act** and **event**. Many underspecified semantic types can get even more complex than this, as illustrated by the hierarchy in Figure 4.10. This hierarchy displays the structure of the underspecified semantic types '`psy`', '`pya`', '`pos`' and '`pas`', that all center around the domains of **possession** and

psychological_feature, but also include the domains of **abstraction** (including the subdomains of **measure, relation** and **phenomenon**), **state, event** and **act**.

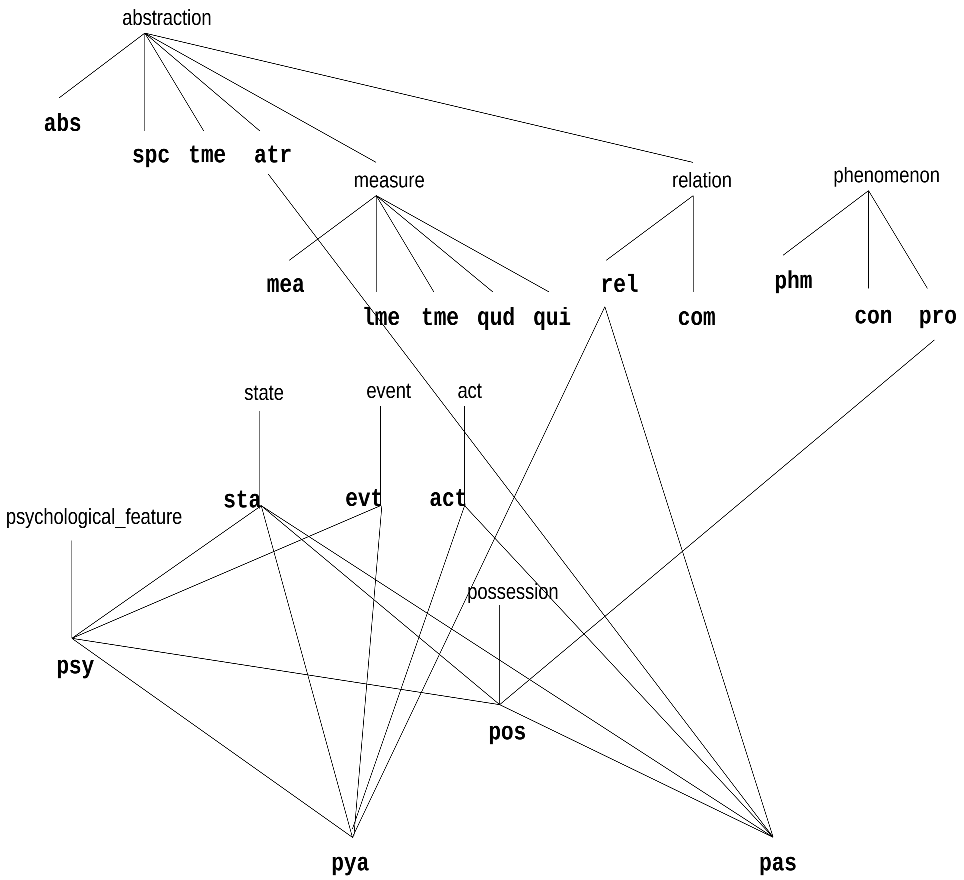

Figure 4.10: Underspecified semantic types in the **possession** and **psychological_feature** domains

Figure 4.11 gives a close-up of the instances of type `'pas'`, corresponding to four systematic polysemous classes.

`act atr pos`	aid allowance inheritance penalty
`atr pos`	asset birthright bounty cost equity gain largess mortmain profit spirituality toll
`atr pos rel`	mileage rate
`atr pos sta`	currency liability wealth

Figure 4.11: Polysemous classes and instances for the type: `'pas'`

Given the WordNet sense assignments, as expressed by the different systematic polysemous classes, there are two basic types (**attribute** and **possession**) central to this collection of nouns. Others enter only through one of the sytematic polysemous classes. The type **act** is given only to the nouns 'aid, allowance, inheritance, penalty', **relation** only to 'mileage, rate' and **state** to 'currency, liability, wealth'. One could easily argue, however, that there is an **act** interpretation also to 'profit, cost' and 'gain', or a **relation** interpretation also to 'aid, allowance, inheritance and liability'.

For further illustration of the underspecification of nouns of this type, consider sentences (23a) through (23c), again from the BROWN corpus, that include underspecified usage of the nouns 'aid, inheritance' and 'profit', covering most or all of the interpretations just mentioned[9].

(23a) The annual spring meeting has given an impetus in three main directions: more, deeper, and more timely political consultation within the alliance, the use of the Organization for Economic Cooperation and Development (when ratified) as a method of coordinating aid to the underdeveloped countries, and the need for strengthening conventional forces as well as the maintenance of the nuclear deterrent.

(23b) The severe yet harmonious vases of the previous fifty years, the Strong Geometric style of the late ninth century, display as firm a mastery of the principles underlying Geometric pottery; but artists now were ready to refine and elaborate their inheritance.

(23c) The company disposed of the pelts, but with what profit, the records do not show.

Underspecified semantic types between the *entity* and *abstraction* domains

The previous examples are still largely restricted to a number of 'abstract' domains. There are however also classes of systematic polysemous nouns that cross between the *abstract* and the *concrete*. A clear example of this are nouns like book that refer to both a **physical_object** (⇒ **entity**) and its **content** (⇒ **abstraction**). There are a number of underspecified semantic types describing such nouns as displayed in Figure 4.12.

[9] Compare recent, corpus-based research on definite descriptions [Poesio and Vieira, 1998], in which it is observed that in many cases human annotators cannot decide between interpretations for certain discourse referents.

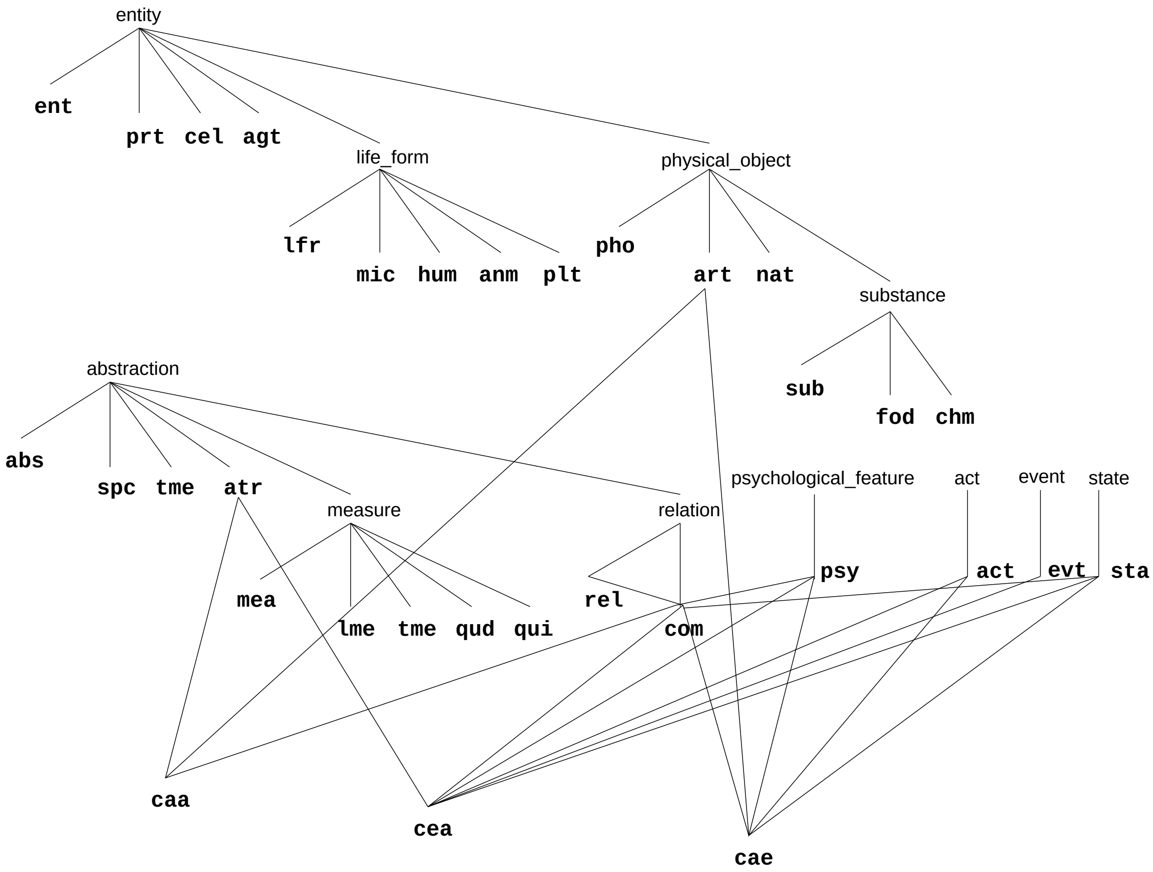

Figure 4.12: Underspecified semantic types between the **entity** and **abstraction** domains

This hierarchy gives the structure of the following underspecified semantic types: `caa`, `cea`, `com` and `cae`. All of these center around the (sub-)domain of **communication**. The types `caa` and `cae` however also reach into the **entity** domain.

The type `'caa'` defines nouns that refer to objects of class **communication** and **attribute**, but are also of the class **artifact** (⇒ **physical_object** ⇒ **entity**).

The type `'cae'` defines nouns referring to objects of **communication**, **psychological_feature**, **act**, **event**, **state** and **artifact**. The following are (select) instances for the seven systematic polysemous classes that correspond to this type:

act art com	documentation drawing recording sketch
act art com psy	art depiction imitation portrayal
act art psy	design music simulation
act art psy sta	abstraction
art com	article bible book catalogue chart journal
art com psy	illustration programme
art com psy sta	picture

Figure 4.13: Polysemous classes and (select) instances for the type: `'cae'`

In the following examples from BROWN, the underspecification of some of these nouns is illustrated. In (24a), 'recordings' refers in equal measure to the object in the hands of the researcher as to its content and even to the act of recording that preceded it. The noun 'catalogue' in (24b) is especially underspecified for the act of assembling instances and the object that results from this. Finally, the underspecified use of 'picture' in (24c), between the concept of a picture and its physical version, is almost idiomatic.

(24a) He had to study the recordings of the language of the chief nation of Ozagen, Siddo.

(24b) Any reputable French interne can supply you with a dozen similar instances, and I'll presently recount a case out of my own personal experience, but, for the moment, let's resume our catalogue.

(24c) The navy captain disclosed also that a list of questions found in Miss Gee's purse would, if completed and handed back, have given the Kremlin a complete picture "of our current anti-submarine effort and would have shown what we are doing in research and development for the future".

Underspecified semantic types within the *entity* domain

Within the entity domain we can also find a large number of systematic relations between classes of nouns. Figure 4.14 gives some examples here, including the types `anf` (**animal**, **food**) and **plf** (**plant**, **food**). Note that the (basic) type **food** itself is underspecified, covering objects that are of class: **food**, **substance**, **form**, **state**, **act** and **natural_object**.

The hierarchy further includes the types `ann` and `pan`, both defining objects that are a **natural_object** (**animal** and **plant**) turned into an **artifact**. This includes nouns like: beaver, chinchilla, mink, alligator, ermine, lapin, raccoon and cotton, hemp, tobacco, calabash, bamboo, kapok, pine, willow.

Additionally, the hierarchy gives the structure for the types `cha` (**chemical** objects with an aspect of **substance** and used as an **artifact**) and `sas` (objects of **substance** with an aspect of **state** and **act**).

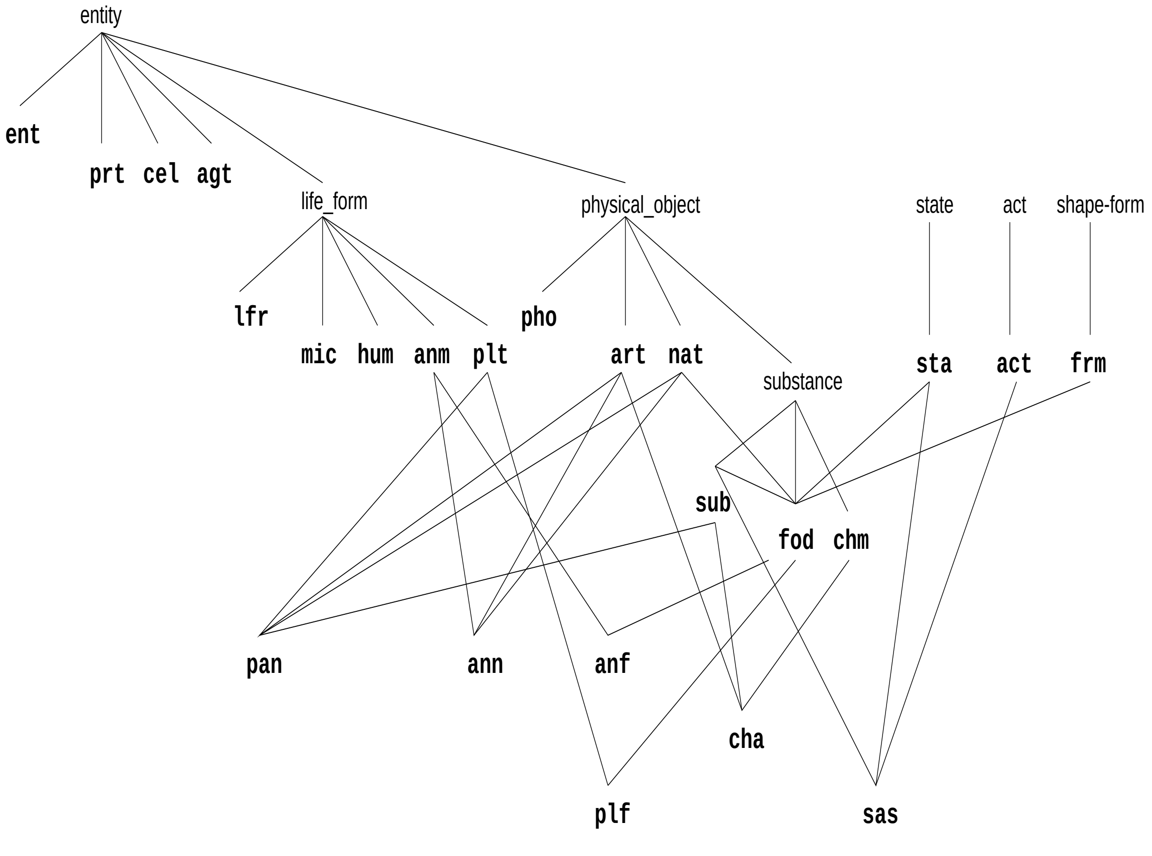

Figure 4.14: Underspecified semantic types within the **entity** domain

The type `anf` only covers the systematic polysemous class `'anm fod'` as shown in Figure 4.15. A case could be made for including also every instance of the class `'anm'`, because in principle every animal could be eaten. This is a question of how *generative* the lexicon should be and if we allow *overgeneration* of semantic objects.

`anm fod`	albacore alewife anchovy blowfish bluefin bluefish bluepoint bonito bream brisling buffalo buffalofish butterfish capon carp catfish cherrystone chow cisco cockle codfish coho cohoe crappie crawdad crawfish crayfish croaker cusk dolphinfish duckling eel elver flatfish flounder fowl grouper grub haddock hake halibut hen herring kingfish lamb langouste limpet lingcod littleneck lobster mackerel mahimahi monkfish mullet muskellunge mussel octopus panfish partridge pheasant pickerel pigeon pilchard plaice pollack pollock pompano porgy poultry prawn puffer pullet quahaug quahog quail redfish rockfish rosefish sailfish saki sardine scallop schrod scollop scrod shad shellfish smelt snail sockeye sprat squid striper sunfish swordfish trout tunny turbot weakfish whelk whitebait whitefish whiting winkle

Figure 4.15: Instances for the type: `anf`

AGENTIVE NOMINALS: **Between the *animate* and *inanimate* domains**

Another example within the **entity** domain crosses between the *animate* (**life_form**) and *inanimate* (**physical_object**). *Agentive nominals* are nouns that refer to both **human** objects with certain occupations and **artifact** objects that came to fullfil those same activities. Examples are printer, computer, trainer, washer (see [Busa, 1997] for an analysis of this group of nouns).

It has to be noted that most of these nouns will not easily occur in discourse with their two related interpretations being expressed at the same time, which warrants a representation using the 'open' dotted type. A 'printer' will be mostly either a human being working on a printing machine, or a machine printing on its own or by command. This is illustrated by the following four sentences in BROWN that include this noun:

(25a) The young printer's apprentice achieved greatness in a half-dozen different fields, as editor and publisher, scientist, inventor, philanthropist and statesman.

(25b) But because the governor was determined that friendship should not influence him one way or the other, he looked for a printer with a knowledge of the law (which Woodruff did not have), and awarded the contract to a lawyer named John Steele who had started a newspaper in Helena the year before.

(25c) It is but justice to Mr. Steele for us to add that, in the above remarks, nothing is intended to his disparagement, either as a lawyer or as a printer.

(25d) Both Alfred Harcourt and Donald Brace had written him enthusiastic praise of Elmer Gantry (any changes could be made in proof, which was already coming from the printer) and they had ordered 140,000 copies - the largest first printing of any book in history.

All instances for the type `'arh'`, covering the class `'art hum'`, are given in Figure 4.16. However, not all instances in this class fit the `arh` type. A number of instances are polysemous in different ways (doll, doormat, gasbag, secretary and the *'art-artist coercion'*: beethoven, brahms, chopin, haydn, stravinsky, wagner) or are more accurately defined as homonyms (boss, coach, dyke, john) with no systematic relation between the **artifact** and **human** interpretations.

`art hum`	accessory accumulator adam antique attache automaton bach backscratcher baggage beethoven blimp boater booster boss bowler brahms brick broadcaster calculator camper chatelaine chopin classic cleaner clotheshorse coach comforter comic commuter compiler computer controller corsair cottar cotter crank cultivator cutter damascene derelict destroyer developer digger diner distributor doll doormat dowser drawer dresser dumbbell dummy dyke edger editor excavator extra fag faggot fagot fake fan fighter figurehead finder gasbag generator gillie goffer governor granny grappler graybeard guy handel harvester haydn holder hood hoodoo hulk hunter incendiary interpreter jean john kingpin knight loafer mannequin mantrap marquise masher master miller mouthpiece mozart mufti nonsmoker pacifier peacekeeper pinhead pirate pivot plunger pouf printer privateer processor prompter puppet pusher putter ramrod reader receiver recorder regulator rep riff rigger riveter rivetter rocker ruler sailor secretary server sewer sham sharpie sharpshooter shaver shooter skirt skivvy speaker spectator spinner spoiler squawker standby starter stravinsky striker stroller supervisor tanker teaser timekeeper tormentor trainer translator transmitter tuner turncock turner vamp viewer voluntary wagner walker washer weekender whaler

Figure 4.16: Instances for the type: `arh`

An overview of all underspecified semantic types in CoreLex

On a whole, CoreLex covers 39,937 noun instances spread over 126 types, of which this section presents a number of examples. For a complete overview with all types, classes and noun instances see the CoreLex webpage (`http://www.cs.brandeis.edu/ paulb/CoreLex/overview.html`) or Appendix A, which gives a list of all 126 types.

4.3.3 Qualia Structures

The underspecified semantic type `acr` binds together five basic types as discussed above: **act, attribute, event, relation, state**. We can further specify the relations between these basic types through a knowledge representation like qualia structure as discussed in Chapter 5.

Qualia Roles

Consider a possible knowledge representation for the type `acr` in Figure 4.17. The representation takes the form of a qualia structure and introduces a number of objects that are of a certain type.

$$
\begin{bmatrix}
\textbf{acr} \\[1em]
\textsc{formal} \;=\; \mathbf{X_1{:}act{\bullet}attribute{\bullet}event{\bullet}relation{\bullet}state} \\[0.5em]
\textsc{constitutive} \;=\; \begin{bmatrix} \textsc{modifier} \;=\; \mathbf{X_1{:}act{\bullet}attribute{\bullet}event{\bullet}relation{\bullet}state(Y_1{:}TYPE)} \\ \textsc{modification} \;=\; \mathbf{Y_2{:}TYPE(X_1{:}act{\bullet}attribute{\bullet}event{\bullet}relation{\bullet}state)} \end{bmatrix} \\[1em]
\textsc{telic} \;=\; \begin{bmatrix} \mathbf{P_1{:}event} \\[1em] \begin{bmatrix} \mathbf{X_2{:}act(R_1{:}TYPE)} \end{bmatrix} \wedge \begin{bmatrix} \mathbf{X_3{:}relation(R_2{:}TYPE,\ R_3{:}TYPE)} \end{bmatrix} \end{bmatrix}
\end{bmatrix}
$$

Figure 4.17: Qualia structure for type `acr`

The FORMAL role introduces an object X_1 of type **act•attribute•event•relation•state**. The CONSTITUTIVE introduces objects that are in a modification relationship with X_1. These can be of a wide range of types, depending on the particular domain. The TELIC expresses a 'deconstruction' of how an event P_1 can be associated with an object of type **act•attribute•event•relation•state**. For instance, the event of 'increase' as in 'increasing the coordination between member states' implies 'increasing' both the **act** of coordinating an object R_1 and the coordination **relation** between two objects R_2 and R_3. All these objects are introduced on the semantic level and correspond to a number of lexical items that will be realized in syntax.

Dotted Types

The type `acr` describes interpretations that can not be separated from each other (the **act** and **relation** aspects are intimately connected). The following representation for type **animal∘food** in Figure 4.18 describes interpretations that can not occur simultaneously but are however related. (Compare the *animal_grinding* lexical rule in Section 2.4.3.) It therefore uses a '∘' instead of a '•' as a type constructor. As explained in Section 3.2.3, the use of the '∘' indicates related meaning aspects that cannot occur simultaneously in discourse, but are still connected. In this case, for instance, the AGENTIVE of the `anf` type represents knowledge about the origin of an object of this type, which is in both classes **animal** and **food**. The first class implies an origin in the event of 'being born', whereas the second class implies an origin through some *grinding* event P_2 that connects objects in both classes. The TELIC role of type `anf` represents the knowledge that objects of this type fulfill different thematic roles according to their specific interpretation in a particular discourse segment. Both sides of the underspecified semantic type (**animal** and **food**) will remain available however, throughout the whole discourse.

$$
\begin{bmatrix}
\textbf{anf} \\
\textsc{formal} = \mathbf{X_1{:}animal{\circ}food} \\
\textsc{constitutive} = \begin{bmatrix} \textsc{modifier} = \mathbf{X_1{:}animal{\circ}food(Y{:}TYPE)} \\ \textsc{modification} = \mathbf{Y{:}TYPE(X_1{:}animal{\circ}food)} \end{bmatrix} \\
\textsc{agentive} = \begin{bmatrix} \mathbf{P_1{:}event} \\ \textsc{theme} = \mathbf{X_2{:}animal} \end{bmatrix} \vee \begin{bmatrix} \mathbf{P_2{:}event} \\ \textsc{theme} = \mathbf{X_3{:}food} \\ \textsc{source} = \mathbf{X_2{:}animal} \end{bmatrix} \\
\textsc{telic} = \begin{bmatrix} \mathbf{P_3{:}event} \\ \textsc{agent} = \mathbf{X_2{:}animal} \end{bmatrix} \vee \begin{bmatrix} \mathbf{P_4{:}event} \\ \textsc{theme} = \mathbf{X_2{:}animal} \end{bmatrix} \vee \begin{bmatrix} \mathbf{P_5{:}event} \\ \textsc{theme} = \mathbf{X_3{:}food} \end{bmatrix}
\end{bmatrix}
$$

Figure 4.18: Qualia structure for type `anf`

Constructing Qualia Structures

Defining the semantic structure (qualia structure) of a class of lexical items is not a trivial task. The examples given above are arrived at by mere introspection. A proper account of underspecified semantic type definitions should be informed by both an 'appropriate' theory of lexical semantics and empirical data. In the next chapter I will show how qualia structures can be generated partly from a corpus based analysis, which maps syntactic structures onto a number of semantic relations between (classes) of lexical items. This, for instance, gives some indication of the type of modifiers that occur frequently with a particular class.

A harder task is to determine the nature of the kind of 'dot' relation between the constituent basic types of an underspecified semantic type. Here we can only arrive at a proper analysis by investigating closely how the objects of such types occur in discourse. That is, if one or more 'sides' of a dotted type occur in the same discourse and thus refer to a *complex object* (as with the `acr` type) or rather to one or more related *simple objects* (as with objects of type `anf`).

4.4 Homonyms

CoreLex is designed around the idea of systematic polysemous classes, excluding homonyms. Traditionally a lot of research in lexical semantics has been occupied with the problem of ambiguity in homonyms. Our research shows however that homonyms only make up a fraction of the whole of the lexicon of a language. Out of the 42,116 noun stems that were derived from WordNet 2,179 are to be viewed as true homonyms because they have two or more *unrelated* senses, less than 5.2%. The remaining 94.8% are nouns that do have (an indefinite number of) different interpretations, but all of these are somehow related and should be inferred from a common knowledge representation. These numbers suggest a stronger emphasis in research on systematic polysemy and less on homonyms, an approach that is advocated here (see also [Kilgarriff, 1992]).

In CoreLex, homonyms are simply assigned two or more underspecified semantic types, that need to be disambiguated in a traditional way. There is however an added value also here because each disambiguated type is underspecified by itself, generating a number of related interpretations. For instance, the polysemous class `'art grs nat pos'` (with member: `bank`) corresponds to a number of underspecifed semantic types: `gsa` (`'art grs'`), `naa` (`'art nat'`) and `psa` (`'art pos'`). We can use a classification algorithm to disambiguate between these by examining the linguistic contexts (surrounding verbs, adjectives, etc.) that are specific to each type. In Chapter 6 I will discuss such an algorithm that can be used both to determine a 'local' interpretation for polysemous lexical items and to disambiguate homonyms.

Chapter 5

Qualia Structure Acquisition

5.1 Introduction

The ontology that was described in the previous chapter groups together nouns in a number of classes that are defined by a set of corresponding intensional objects, or types. Each type is either basic or underspecified. Underspecified semantic types are organized as a combination of basic types, using two type combinators: the open ('$\circ$') and closed ('$\bullet$') dot. Such information on the semantic type of a lexical item is at the core of a representation of its lexical semantic or *qualia structure*.

Next to this, there should be a representation of how lexical items of this type show up with lexical items of other types. This can be analysed by looking at the *functional structure* of sentences in which these lexical items occur. Here, I take this to be an analysis of how semantic objects that correspond to such lexical items are modified and how they show up as actors or modifiers themselves.

For instance, a noun can be modified by an adjective, or some other noun. Also, it can be the subject, direct object or adjunct of some verb. At this level of analysis we arrive at the surface of language, its *syntactic structure*.

So we work our way back from syntactic structure on the surface, through functional structure on the semantic level, to qualia structure as an abstraction over the functional structure of a particular lexical item or class of lexical items.

The idea of linking syntactic (surface) to semantic (deep) structure is of course a very old one, going back to the early days of transformational-generative grammar and beyond that to (European) structuralism and functionalism[1]. Taking a next step of accumulating semantic structure into lexical semantic definitions is something not often encountered in linguistic theory. In AI, NLP and adjacent fields however, this is not uncommon practice in order to learn something about a specific domain. The fields of *knowledge acquisition* and *machine learning* produced a wide variety of methods to acquire new, or update existing knowledge, often from text. The task described in this (and the next) chapter is to combine similar methods with a traditional linguistic analysis in order to acquire lexical semantic knowledge representations.

Such an approach has in fact been taken by a number of researchers, see for instance the collections of papers in [Zernik, 1991] and [Boguraev and Pustejovsky, 1996]. Much of this research however is devoted to the extraction of lexical semantic structure for verbs, rather than nouns, which is the topic of this thesis. On the other hand, if researchers address the issue of automatic lexical knowledge extraction for (proper) nouns, then often this is done in the context of *information extraction*, a field of study that tries to obtain very specific information in large bodies of text (see e.g.: [Hobbs *et al.*, 1997]). The objective here is somewhat different. Information extraction systems mostly deal with a rather limited set of database relations that need to be 'populated' by examining texts for relevant occurrences. In practice, this constitutes mostly a largely restricted ontology of classes and relations between these classes for some specific small domain.

The research described here deals with a full ontology of semantic classes, applied to several different domains. This gives one the opportunity to compare classes of lexical items in terms of their different linguistic behaviors. Classes will occur more frequently with some classes and less with others, depending also on the domain we are looking at. Some of these experimental results will be discussed in Section 5.4, but first we have to take a closer look at how such knowledge is acquired by analysing the syntactic and functional structure of natural language text.

[1] According to [Luelsdorff, 1994], '[...] Tesnière's treatment of the structure of the sentence, which consistently distinguished between two levels of sentence patterns [...], has been well known in Prague since the 1930's.'

5.2 Syntactic Structure

This section describes in detail which phrase structures are recognized. First, however I discuss some of the formal background of syntactic structure in natural language and the inherent restrictions of *regular expressions* that are used to recognize it.

5.2.1 Recursion and Formal Grammar

To recognize the syntactic structures in a given corpus we would like to be able to use a recursive grammar like the one in Figure 5.1 that defines the structure of a prepositional and nominal phrase. It includes recursion over NP's and PP's, by which we can describe phrases like 'the cat on the mat by the door of the house in the street ...'.

PP	**:**	**P NP**
NP	**:**	**Det N**
NP	**:**	**Det N PP**

Figure 5.1: A simple context free grammar for PP's and NP's

In order to capture the recursive structures that can be generated by this grammar (and as they actually occur in natural language) we need to 'flatten' the grammar to make it fit a description with regular expressions. Normally we cannot do this with a context free grammar (CFG), but in this case however, we can, because the grammar in Figure 5.1 is of a particular subset of context free grammars (left-linear grammars) that are regular [Partee *et al.*, 1993]. The flat version of this same grammar then corresponds to the following regular expressions:

NP	**:**	**DET N (P DET N)***
PP	**:**	**(P DET N)$^+$**

Figure 5.2: Regular expressions for PP's and NP's

5.2.2 Phrase Structure

The pattern matcher consistes of a number of regular expressions that are matched against an (English) text that has been run through a cascade of pre-processing programs. In sequence, these include *tokenizing*, *part-of-speech tagging* and *stemming* the text:

Tokenization

The tokenizer used in this research was constructed at Brandeis as a general tool in natural language analysis, not specifically tuned to one or more specific document forms [Verhagen, 1996]. It splits texts into sentences, one per line, using punctuation marks ('.', '?' , '!') as split-off points. Abbreviations, numbers and other usage of dots ('.') are preserved however, that is, they are not treated as split-off points. All other word-internal punctuation like hyphens ('-' , '_'), symbols ('$' , '%'), and others are preserved also.

Part-of-speech tagging

The tokenized version of a text is part-of-speech tagged with an off-the-shelf tagger [Brill, 1992]. The tagger is not trained for a specific document, or set of documents. I assume throughout the research that the default training of the BRILL tagger on the BROWN corpus is general enough to handle most kinds of text.

The accuracy of the BRILL tagger is assumed to be around 96% to 97%. To test this, a non-representative analysis was performed on the STOCKS text, a one-page document of 740 words about the performance of the Wall Street stock exchange. It was acquired through the YAHOO online news service. Before tagging, the text was stripped of 'HTML' tags and tokenized with the program mentioned before, which produced 846 tokens to be tagged. Of these 33, or 3.9%, were judged to be wrong as illustrated in Figure 5.3. A number that roughly agrees with the accuracy rate just mentioned. Note however, that the most common error is to tag a quotation marker as a noun (at the beginning of a sentence). This, in fact is a problem that is connected to some restrictions on proper tokenization rather than part of speech tagging.

Frequency	Actual Tag	Proper Tag	Comments
8	NN	‘	quotation mark as noun
4	JJ	NN	noun as adjective
3	NNP	NN	noun as proper noun
2	NNPS	NNS	noun plural as proper noun plural
2	IN	RB	adverb as preposition
2	NN	&	& as noun
2	NN	RB	adverb as noun
1	IN	WDT	‘`that`’ as preposition
1	JJ	RB	adverb as adjective
1	NN	JJ	adjective as noun
1	NN	CD	number as noun
1	NNP	‘	quotation mark as proper noun
1	RB	NN	noun as adverb
1	RB	VB	verb as adverb
1	VB	CD	number as verb
1	VBG	JJ	adjective as gerund
1	VBG	‘	quotation mark as gerund

Figure 5.3: Errors in BRILL part-of-speech tagging of the STOCKS text

Stemming

The final stage of preprocessing is to stem each token in the text. Here I use a stemmer that was developed at Brandeis [Verhagen, 1996] by enhancing a stemming algorithm widely used in information retrieval [Porter, 1980], with the CELEX[2] database on English morphology. Stemming is trivial for most tokens, like symbols, determiners and prepositions, which are ‘stemmed’ by substituting the same token, but rather intricate for especially nouns and verbs. In most cases the algorithm follows the Celex database, unless the particular token (word) doeas not occur there. In that case the algorithm relies on a number of general heuristics of the Porter stemmer.

Recognized Phrase Structures

The regular expressions that are matched against the tokenized, tagged and stemmed version of the text cover three main syntactic structures: complete (S), or incomplete (VP) sentences, including prepositional complements; nominal phrases (NP), including prepositional modifiers; and adjectival

[2]For CELEX, see `http://www.kun.nl/celex/`

phrases within NP's (AP). Within these main groups there are a number of syntactic variants that are described below. The definitions given here are reminiscent of the regular expressions that are used for their implementation in PERL.

To allow for readability of the patterns some abbreviations are employed that are used also in the actual implementation. The following two definitions reduce some complex patterns into one meta-symbol, that is then used in the actual patterns below. The first definiton gives a complex expression corresponding to an **NP**, while the second defines a 'verb group' or **VG**.

NP	**:**	**PREDET* (DET‖WHDET)* NUM* ADJ* NN* HEAD_NN**
VG	**:**	**AUXVERB* ADVERB* VERB PGP***

Figure 5.4: Definitions for nominal phrase and verb group

In fact, one other meta-symbol is used: **NON_NN**. This matches anything that is not a *noun* or a *proper name* and is used where ever an NP is to be matched at the end of a pattern, in order to ensure the NP also really ends there.

Given these definitions, the recognized phrase structures are then as follows:

1. S/VP

(a) **(PRONOUN‖NP) VG NP (PREP NP NON_NN)?**

Description : PREPOSITIONAL ADJUNCT

Example : 500 stocks fell 1.43 points to 628.37

(b) **(PRONOUN‖NP) VG PREP NP (PREP NP NON_NN)?**

Description : PREPOSITIONAL ARGUMENT (particle)

Example : stocks collapsed in early afternoon on shaky investor confidence

(c) **TO VG NP NON_NN**

Description : INFINITIVAL

Example : to buy battered shares

(d) **NP MODAL? HAVE? BE ADVERB* PSP ADVERB* (BY NP NON_NN)?**

Description : PASSIVE

Example : the market was boosted initially

(e) **VG PREP VG PREP NP NON_NN**

Description : VERBAL COMPLEMENT ??

Example : concerned about looking for a chance

2. NP

(a) **NP PREP NP**

Description : PREPOSITIONAL MODIFIER

Example : massive losses in the heaviest trading day

(b) **NP POS NP**

Description : POSSESSIVE

Example : Tuesday afternoon's resounding bounce

(c) **PREDET* (DET||WHDET)* NUM* ADJ* NN* NN HEAD_NN NON_NN**

Description : NOUN-NOUN MODIFICATION

Example : consumer prices

3. AP

(a) **PREDET* (DET||WHDET)* NUM* ADJ* ADJ NN* HEAD_NN NON_NN**

Description : ADJECTIVE-NOUN MODIFICATION

Example : battered shares

All of these patterns are non-lexical, that is, they describe general sequences of syntactic classes of lexical items (except for the auxiliary and modal verbs in pattern 1.d). However, some lexical restrictions do apply. Two lexical filters ensure that only relevant occurrences of the patterns match. To be precise, one filter checks S/VP patterns for occurrences of auxiliary verbs (to be, to have and to do) as main verbs. In that case the pattern will not be taken in consideration. A second filter checks the **NP Prep NP** pattern against a list of 'relevant' prepositions. This is needed, mainly because the part-of-speech tagger does not always assign the preposition tag ('IN') correctly. It does assign IN for instance sometimes to the conjunction 'that'.

Next to these filters, two additional patterns also use lexical items. These patterns recognize *partitive* constructions, where two nouns express objects that are in part-whole relation to each other. The patterns are in fact variations on two patterns given earlier:

1. **NP (have||contain||include) NP NON_NN**

   Description : PARTITIVE VERB

   Example : computer has a clock

2. **NP of NP**

   Description : PARTITIVE OF

   Example : the lower-left section of the dialog box

Although lexically constrained, these two patterns are still very general. Only a small percentage of the occurrences that match them will indeed express a partitive construction. We therefore need to apply a semantic filter that only considers likely combinations of nouns and disregards the rest. I will get back to this in the next section on the analysis of functional structure (Section 5.3.2).

Unrecognized Phrase Structures

All the patterns given here are the ones that are actually recognized, but they are of course by no means the only patterns to be found. There are large numbers of different patterns that occur in actual text, but most of these are not general enough to motivate an additional rule that describes them. In theory we could include every possible pattern in order to have complete recall. This however would be quite infeasible in practice and also not desirable, because the patterns would then rely too heavily on a specific domain, a particular kind of text, or even on the writing style of a particular author. Instead I chose to design the pattern matcher as general as possible in order to be able to handle a large variety of texts from different sources and domains.

Corpora used

The corpora that were used in the research described in this thesis are given in Figure 5.5. Only one small text was used in constructing the pattern matcher (STOCKS), while the classification algorithm, described in Chapter 6, was constructed using a number of small and medium sized texts (especially STOCKS, ORWELL, MR and DARWIN). Some larger corpora (BROWN, WSJ and PDGF) were used as a test set. Some other smaller texts were included specifically to evaluate the effects of different styles and genres on both pattern matching and classification, for instance NREN and ULYS are specific in respect to *bureaucratic* and *literary* language use.

WSJ	:	*Wall Street Journal* 198? issue (first 50,000 lines)
BROWN	:	Brown corpus
PDGF	:	2116 abstracts on Platelet Derived Growth Factor (obtained through MEDLINE)
DARWIN	:	*Origins of Species* by Charles Darwin
MR	:	*MacRef user manual* for Apple Macintosh computers
ULYS	:	First chapter of *Ulysses* by James Joyce
ORWELL	:	Essay by George Orwell on *Politics and the English Language*
NREN	:	U.S. government document on the *National Research and Education Network*
STOCKS	:	News article on stock market performance (obtained through YAHOO)

Figure 5.5: Corpora used

Accuracy of syntactic analysis

Precision and recall on the pattern matcher can be measured in two ways. First, we can measure recall by computing the percentage of nouns that are covered by the patterns that are matched. Precision (the number of the nouns that are covered that are actually nouns) is much harder to measure, but depends for a large part on the accuracy of the output of the part-of-speech tagger. Here I assume the earlier given estimate of 96% to 97%. Recall varies from 48% to 82%, depending on the nature of the text, see Figure 5.6. The pattern matcher was constructed on the basis of a small business text (STOCKS) about stock market performance and includes straightforward general patterns that reflect rather 'non-creative' language only. Consequently, the matcher performs better on technical texts, like WSJ and MR, than literary ones, like ULYS and ORWELL. The good performance of the pattern matcher on the PDGF corpus is somewhat surprising, because the domain (biomedical abstracts) is fairly different from the business domain of the STOCKS text, although also technical in nature.

corpus	*tokens*	*noun stems*	*matched*	*recall*
WSJ	1208408	10984	8621	78.49
BROWN	1156740	15618	11546	73.93
PDGF	540636	4606	3626	78.72
DARWIN	168619	2370	1663	70.17
MR	23059	493	403	81.74
ULYS	8536	718	345	48.05
ORWELL	6177	577	315	54.59
NREN	1410	146	88	60.27
STOCKS	846	85	62	72.94

Figure 5.6: Recall on the pattern matcher

These numbers on precision and recall give some indication on the performance of the pattern matcher in terms of the individual nouns matched. However, to measure the accuracy of the pattern matcher in terms of the actual patterns as a whole we need a different approach. It would be quite complex and tedious to check every pattern manually and thereby give an exact measure of the accuracy of the patterns that were matched, but we can get some idea of the accuracy into some extent by doing an experiment that employs a hand tagged and parsed corpus like the PENN TREEBANK. Consider the following first 10 sentences from the PENN TREEBANK version of the Wall Street Journal:

(26a) Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

(26b) Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

(26c) Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC, was named a nonexecutive director of this British industrial conglomerate.

(26d) 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.

(26e) The asbestos fiber, crocidolite, is unusually resilient once it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said.

(26f) Lorillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.

(26g) Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

(26h) A Lorillard spokewoman said, "This is an old story.

(26i) We're talking about years ago before anyone heard of asbestos having any questionable properties.

(26j) There is no asbestos in our products now." Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes.

According to the PENN TREEBANK[3] these sentences contain the following NP-Subjects (**NP-SBJ**), some of them occurring in the same sentence.

(27a)	[NP-SBJ[NP Pierre/NNP Vinken/NNP],/,[ADJP[NP 61/CD years/NNS]old/JJ],/,]
(27b)	[NP-SBJ Mr./NNP Vinken/NNP]
(27c)	[NP-SBJ-1[NP Rudolph/NNP Agnew/NNP],/,[UCP[ADJP[NP 55/CD years/NNS]old/JJ]and/CC[NP[NP former/JJ chairman/NN][PP of/IN[NP Consolidated/NNP Gold/NNP Fields/NNP PLC/NNP]]]],/,]
(27d)	[NP-SBJ[NP[NP A/DT form/NN][PP of/IN[NP asbestos/NN]]]
(27e)	[NP-SBJ researchers/NNS]
(27f)	[NP-SBJ[NP The/DT asbestos/NN fiber/NN],/,[NP crocidolite/NN],/,]
(27g)	[NP-SBJ it/PRP]
(27h)	[NP-SBJ[NP even/RB brief/JJ exposures/NNS][PP to/TO[NP it/PRP]]]
(27i)	[NP-SBJ researchers/NNS]
(27j)	[NP-SBJ[NP Lorillard/NNP Inc./NNP],/,[NP[NP the/DT unit/NN][PP of/IN[NP[ADJP New/JJ York-based/JJ]Loews/NNP Corp./NNP]][SBAR[WHNP-2 that/WDT][S[NP-SBJ *T*-2/-NONE-][VP makes/VBZ[NP Kent/NNP cigarettes/NNS]]]]],/,]
(27k)	[NP-SBJ-2 preliminary/JJ findings/NNS]
(27l)	[NP-SBJ the/DT latest/JJS results/NNS]
(27m)	[NP-SBJ A/DT Lorillard/NNP spokewoman/NN]
(27n)	[NP-SBJ This/DT]
(27o)	[NP-SBJ We/PRP]
(27p)	[NP-SBJ anyone/NN]
(27q)	[NP-SBJ asbestos/NN]
(27r)	[NP-SBJ There/EX]
(27s)	[NP-SBJ Neither/DT[NP Lorillard/NNP]nor/CC[NP[NP the/DT researchers/NNS][SBAR[WHNP-3 who/WP][S[NP-SBJ *T*-3/-NONE-][VP studied/VBD[NP the/DT workers/NNS]]]]]]

Of the 19 NP-Subjects only 4 ((27g), (27k), (27l), (27p)) were recognized by the pattern matcher. The mismatch on the other 15 has a variety of reasons. A number of the subjects ((27a), (27b)) are headed by a proper noun (`NNP`), instead of a noun (`NN` or `NNS`) and are therefore disregarded. Others ((27n), (27r)) are headed by lexical items of different syntactic categories, that are equally disregarded. Still some other subjects are just to complex to be handled by the patterns described above ((27c), (27f), (27h), (27j), (27s)). Finally, the remaining subjects that were missed occur in patterns that were not general enough to be included in the pattern matcher. Additionally, the pattern matcher assigned one pronoun ('it' in (27h)) inaccurately the status of subject.

[3]The presentation of syntactic bracketing as presented here is an adaptation of the original PENN TREEBANK format.

These results lead to a recall measure of roughly 20% (4 out of 19) and a precision of 80% (4 out of 5, the only mistake is 'it' in (27h)) . Precision is thus quite good, but recall rather low. However, a lot of the reasons for this low recall can be easily fixed. We would need to include proper nouns as potential subjects, as well as some other syntactic categories. This would raise the recall to about 40% (8 out of 19). Additionally, we would need to expand the coverage by including more complex and 'exotic' patterns. On a whole, however, low recall is a less serious problem, because it is easier to fix, then low precision.

5.3 Functional Structure

The syntactic analysis just described goes hand in hand with a 'deeper' semantic analysis that gives some indication of the *function* of the phrases that have been extracted. This process employs the semantic class of the heads of those phrases, in particular the nominal heads of NP's and their syntactic structure in respect to prepositions and verbs. In fact, determining the functional structure of a phrase incorporates an analysis of the interaction between the semantic classes of all heads involved. In the case of a prepositional adjunct to a verb this would include the head of the verb phrase, the nominal head of the embedded NP and the prepositional head.

The analysis I describe here however is somewhat more restricted and includes only the interaction between the nominal head of the embedded NP and the lexical form (not its semantic class) of the prepositional head. This analysis is much less accurate unfortunately than it could have been by including additional information from the semantic classes of both the verbal and the prepositional heads, but this would have required a lot more additional research that was not the immediate focus of this thesis. The analysis does however use the CoreLex database on nouns as described in part in the previous chapter.

The functional structure that is assigned to NP's and their nominal heads takes the form of *thematic roles* in the case of nominal and prepositional arguments and adjuncts to verbs and prepositional adjuncts to nouns. Other aspects of functional structure I consider consist of an analysis of *meronymy*, or part-whole constructions, and nominal and adjectival modification of nouns. In the following

I discuss some of the details of both thematic role assignment and of the analysis of part-whole relationships and modification.

5.3.1 Thematic Roles

The assignment of semantic roles to syntactic phrases has some tradition in linguistic theory, but originates primarily with [Tesniere, 1959] and was extended by [Fillmore, 1968]. The basic idea is to assign an additional ('semantic') structure to a sentence, or sentential clause that overlays the syntactic one. The use for such additional structure is most clear in the case of passive constructions, where the syntactic subject of the phrase is not the logical or semantic subject. In addition, the passive involves a prepositional adjunct (headed by the prepostion 'by') that does include the semantic subject, or rather the *agent* of the event that is expressed by the sentential clause. Such a treatment of prepositional adjuncts (and arguments) can be extended to include a number of functions or *roles* performed by their embedded NP's. Here I consider the following roles:

AGENT	:	entity that initiates the action
BENEFICIARY	:	entity that benefits from the action
CO_AGENT	:	entity that co-initiates the action
DESTINATION	:	entity to which something is moved due to the action
DURATION	:	time span of the action
INSTRUMENT	:	entity that is used in the action
LOCATION	:	specification of the location of the action
SOURCE	:	entity from which something is moved due to the action
THEME	:	entity that undergoes the action
TIME	:	specification of the time of the action
TRAJECTORY	:	entity through which the action is performed

Figure 5.7: Thematic roles used

Unfortunately, the assignment of thematic roles is not at all straightforward. There is an accumulation of ambiguities that leaves quite some room for rather different assignments. First, the definitions of individual thematic roles, as the ones given in Figure 5.7 are rather intuitive and different linguists distinguish different numbers of thematic roles and giving them different interpretations. Second, the assignment of thematic roles depends for some part on the kind of preposition involved. Unfortunately, prepositions are among the most ambiguous words in any language and therefore not completely reliable as semantic indicators. For instance, the preposition 'for' indicates DURATION if a temporal noun is involved ('They talked for hours.'), or BENEFICIARY otherwise ('This present is for our daughter.').

The analysis described here cannot remedy the first ambiguity, simply because it is inherent to the nature of the problem itself. We first need a very broad and deep analysis of thematic role assignment in order for it to be standardized into some respect, with most linguistic analyses agreeing. The current analysis that is used here is just one, without any general claims, but with a direct use in qualia structure analysis, semantic classification and underspecified semantic tagging as discussed further below.

The second kind of ambiguity can be resolved largely by taking into account the semantic class of the head noun in embedded NP's. The interaction of the prepositional head and the semantic class produces a thematic role for the NP that is disambiguated in respect to the ambiguity of the preposition, although not entirely in respect to the *polysemy* of the noun.

Consider for instance the example of head nouns in BROWN that have been assigned the DURATION role as given in Figure 5.8. This role is given to head nouns of class `'tme'` or `'tmv'` if they are embedded in an PP that is headed by 'for' or 'of'.

Frequency	Head Noun	Semantic Class
24	day	tme
14	month	tme
12	future	tme
12	week	tme
8	while	tme
5	century	tme
4	instant	tme
4	season	tme
3	half	tme
3	summer	tme
3	vacation	tmv
3	weekend	tme
2	decade	tme
2	era	tme

Figure 5.8: Most frequent head nouns in BROWN with DURATION role

In most cases the DURATION role is given to nouns that are of semantic class `'tme'`. In one case however, it is given to a noun (vacation) of class `'tme'`, which is underspecified for the semantic types **event** and **time**. Sentence (28) is one of the instances of vacation that was taken from the BROWN corpus.

(28) But if we closed the store for a vacation, we would lose our customers to the chain stores in the next block.

This example illustrates why it is hard in such a case to distinguish between different polysemic interpretations of the noun in question. This however does not necessarily effect the thematic role assignment, because, like in this case, the 'duration' aspect is expressed by both of the systematically related interpretations (the 'time period' and the 'event' of vacation).

5.3.2 Modification and Part-Whole Information

Next to thematic roles I consider the semantic relations of *modification* and *part-whole*. Both of these are not traditionally considered to be represented through thematic roles and I therefore discuss them here separately, although they are not treated differently from thematic roles in the actual implementation.

Modification

Modification as discussed here occurs within the NP, through either *adjectival* or *nominal* modification. In both cases their appropriate semantic (functional) analysis depends largely on a resolution of the syntactic ambiguities that are frequent in these structures. Consider for instance (29):

(29) old men and women

Obviously there are two possible syntactic analyses, leading to important different functional structures. In analysis (30a) both men and women are old, whereas in (30b) only the men are.

(30a) [old [men and women]]

(30b) [[old men] and women]

In the actual implementation however, this problem will not be encountered, simply because conjunction within NP's is not handled by the available patterns. Nevertheless, similar problems also occur with even simpler constructions like example (31) (from [Johnston *et al.*, 1995]):

(31) sound control panel

Also here there are two possible syntactic analyses, leading to different functional structures. Analysis (32a) assigns (simple) modification over the *compound* noun 'control panel', whereas (32b) assigns (complex) modification over the (simple) noun 'panel'.

(32a) [sound [control panel]]

(32b) [[sound control] panel]

This problem, in fact involves both nominal and adjectival modification. Consider for instance the rather similar example (33), involving the adjective 'chief':

(33) chief investment strategist

This phrase could be analysed again in two ways:

(34a) $[\mathsf{chief}_{adj}\ [\mathsf{investment}_{noun}\ \mathsf{strategist}_{noun}]_{NP}]_{NP}$

(34b) $[[\mathsf{chief}_{adj}\ \mathsf{investment}_{noun}]_{NP}\ \mathsf{strategist}_{noun}]_{NP}$

The corresponding semantic structures would be something like:

(35a) $strategist(x) \wedge chief(x) \wedge investment(y) \wedge relation(x, y)$

(35b) $strategist(x) \wedge chief(y) \wedge investment(y) \wedge relation(x, y)$

The syntactic analysis as employed by our implementation is only based on part-of-speech tags and is therefore not spophisticated enough to resolve any such ambiguities. Instead, a general approach is used that takes modification always over the *head* of the NP, which is always taken to be the rightmost noun. Although this is a drastic oversimplification of the actually occurring language facts, it does work well enough for one of the main purpose of this thesis, which is semantic classification for semantic tagging. On the other hand, the sometimes inaccurate syntactic analyses unfortunately lead to similar inappropriate functional structures that ultimately also corrupt some of the analyses of qualia structure that are discussed in the next section. However, because of the size of the corpora that are used, I hope this corruption at least to be limited.

Part-Whole Information

Like in modification, the analysis of part-whole relationships between nouns depends for some part on the accurate analysis of the syntactic structure of NP's. For a proper assignment it is necessary to know what the nominal heads of both NP's are. Again, however, our analysis depends only on part-of-speech information and is therefore not sophisticated enough for analyzing such syntactic structure accurately. The patterns that are used can only approximate this by using the general heuristic of the head as the rightmost noun. Nevertheless, this seems to be accurate in most cases in English, allowing automatic part-whole analysis as discussed in this section.

Unlike modification however, part-whole analysis depends also largely on an appropriate semantic analysis of both nouns involved. We can automatically find a considerable number of potential part-whole constructions in a large enough corpus, but not all of these instances are in fact describing such a relation. For instance, the auxiliary verb 'to have' is a good indicator for this relation, but only in some cases. Consider the following example from MR:

(36a) some programs have special installation instructions

(36b) computer has a control panel

(36c) lighting fixtures contain magnets

(36d) computer has a clock

(36e) computer has a built-in floppy disk drive

(36f) many programs have a font menu

(36g) computer has built-in appletalk capability

(36h) network has zones

(36i) computer has file

(36j) all programs have this capability

(36k) computer has a separate monitor

(36l) a new floppy disk has no information

(36m) the printer has paper

(36n) guests have access

(36o) the disk has problems

(36p) many programs have the computer make ...

(36q) some printers have an energy (saving feature)

As shown, there are 17 (different) instances of the pattern '**NP have NP**'. Of these, 11 ((36a) through (36k)) are more or less acceptable as descriptions of part-whole relationships between the nouns involved. Of the remaining constructions, 3 are questionable ((36l) through (36n)) and 2 are wrong ((36o) and (36p)). Example (36q) would have been acceptable if the part-of-speech assignment to 'saving' had been accurate as an adjective instead of as a verb, leading to the more complex NP 'an energy saving feature' with 'feature' instead of 'energy' as the head.

To select only the appropriate constructions a semantic filter is applied that uses the semantic type of the nouns under consideration and some general heuristics to determine if a part-whole relation is likely:

- Two nouns are *not* partitive if the type of any of them is in a list of semantic types that are in general *unlikely* to be used in the expression of a part-whole relationship. This includes for instance **action**, **state** and other **event** types.

- Two nouns are partitive if the type of the second noun is in a list of semantic types that are in general *likely* to be used in the expression of a part-whole relationship. This includes for instance the types **artifact**, **part** and **location**.

- Two nouns are partitive if both are of the same type.

5.4 Qualia Structure

Qualia structure in the context of this thesis is assumed to be a representation of the functional structure of an individual or class of lexical items. Thus, by this definition qualia structure is not something completely different from the functional structure of sentences and clauses, but rather an abstraction over it that presents an image (in practical terms more a 'snapshot') of lexical meaning relative to a particular corpus or domain. Although this approach to qualia structure is somewhat different from the one discussed in [Pustejovsky, 1995], elsewhere a similar approach has been advocated [Pustejovsky *et al.*, 1993], emphasizing also the dependence of qualia structure acquisition on the nature of particular corpora [Johnston *et al.*, 1995].

Here, qualia structure includes three roles: FORMAL, CONSTITUTIVE, AGENTIVE_TELIC. The FORMAL role is a list of the basic types that correspond to the type of the lexical item (or class of lexical items) under consideration. The CONSTITUTIVE role includes a list of (thematic) roles expressing aspects of *modification* by nouns and adjectives, while the AGENTIVE_TELIC role includes a list of thematic roles that are connected to **events** as expressed by verbs. Each (thematic) role includes a list of semantic types that occur with the lexical item(s) under description in this role with their *frequency* and *mutual information* score.

Mutual Information

The influence of frequency (sparseness of data) is smoothed out somehow by using mutual information (MI) scores rather than raw frequencies [Church and Hanks, 1990]. Computing MI scores is by now a standard procedure for measuring the co-occurrence between objects relative to their overall occurrence. We can use this definition to derive an estimate of the *connectedness* between words, in terms of collocations [Smadja, 1993], but also in terms of phrases and grammatical relations [Hindle, 1990]. For instance the co-occurrence of verbs and the heads of their NP objects (N: size of the corpus, i.e. the number of tokens):

$$C_{obj}\ (v\ n)\ =\ log_2\ \frac{\frac{f(v\ n)}{N}}{\frac{f(v)}{N}\ \frac{f(n)}{N}}$$

Cospecs

As discussed in Chapter 3 the representation of qualia structure also involves representing the syntactic patterns (or *cospecs*) that correspond to the functional structures in the qualia roles. Here, cospecs are represented as ordered lists of 'tokens' of the following form: **STEM:POS:SEM**.

- **STEM** : the morphological stem of the lexical item (see Section 5.2.2)
- **POS** : ranges over: N, V, A and P, for noun, verb, adjective and preposition respectively
- **SEM** : CoreLex basic or underspecified semantic type

The cospec representation also includes the *frequency* of the particular syntactic pattern and the head of this construction.

Verbal and adjectival semantic types

Next to the CoreLex nominal semantic types that were discussed in Chapter 4, the qualia structures also use verbal and adjectival semantic types. These, however, have little theoretical value. They were constructed primarily for unity of representation, but are based on a completely automatic analysis of sense distributions in WordNet (using the *lexicographic file number* as 'top sense' indicator) that does not distinguish between homonymy and polysemy and does not acknowledge anything like systematic polysemy.

In summary, these types were constructed by duplicating the analysis for nouns as described in Chapter 4, but without any human interpretation of the data before or after. To extend the accuracy of the qualia structure representations by using appropriate verbal and adjectival types would be important future work, but is beyond the scope of the thesis.

Qualia Structure: An example

Figure 5.9 is an example of a qualia structure representation derived from the PDGF corpus for a particular class of lexical items (`'fre'`). Given the size of the corpus (540,636 tokens), most representations are rather big, including a number of different (thematic) roles that are instantiated by different semantic types and cover a large number of different cospecs. The example here is therefore chosen because of its reasonable size, not because it would be representative for the derived representations in general.

The underspecified semantic type `'fre'` corresponds to four basic types that are represented in the FORMAL role: **act** (`'act'`), **event** (`'evt'`), **form** (`'frm'`), **state** (`'sta'`). The CONSTITUTIVE role gives a probabilistic representation of the types of lexical items that can modify (**adj_mod, theme**) or be modified (**nom_mod_of**) by nouns of type `'fre'`. Each type is followed by its frequency and a mutual information score. Both these numbers are relative to the co-occurrence of this type with type `'fre'` for the given attribute.

For instance, nouns of type `'lac'` (**act, location, location_geographical**) co-occur 5 times in PDGF in a modifying relation with nouns of type `'fre'`. This same information is repeated in the COSPEC part of the representation, but here actual lexical instances and syntactic patterns are represented also. For instance, the 5 co-occurrences of type `'lac'` and `'fre'` turn out to be 5 occurrences of the noun-noun pattern: **fracture:N:fre repair:N:lac**.

The attribute **theme** within the CONSTITUTIVE role represents a functional relation between nouns of type `'prt'` (**part**) and nouns of type `'fre'`. As the corresponding syntactic information in the COSPEC representation shows, this derives from the use of the preposition 'after', connecting the nouns 'bone' and 'fracture'.

There is no information represented in the AGENTIVE_TELIC role. This means that in the PDGF corpus no co-occurrences could be detected between nouns of type `'fre'` and any verb. This, however, depends largely also on the accuracy of the pattern matcher, which only covers some of the more general verb phrase patterns as discussed in Section 5.2.2.

fre

CORPUS = PDGF

QUALIA =

FORMAL = [**act evt frm sta**]

CONSTITUTIVE =

adj_mod
__ 1 7.66
aa 1 5.23

nom_mod_of
lac 5 7.75
pro 3 6.77
aes 1 5.10
psy 1 10.49

theme
prt 1 5.13

AGENTIVE_TELIC = __

COSPEC =

ADJ_MOD =
2 healing [healing:A:aa, fracture:N:fre]
1 nuclear [nuclear:A:aa, protrusion:N:fre]
1 scleral [scleral:A:__, incision:N:fre]

NOM_MOD_OF =
5 repair [fracture:N:fre, repair:N:lac]
3 healing [fracture:N:fre, healing:N:pro]
1 hematoma [fracture:N:fre, hematoma:N:psy]
1 wound [incision:N:fre, wound:N:aes]

THEME = [**1 bone [bone:N:prt, after:P, fracture:N:fre]**]

Figure 5.9: Qualia structure and cospecs for type `fre`, as derived from PDGF

Default types

As mentioned before, by comparing derived qualia structures we can analyse into some respect the differences and agreements in lexical semantic structure between corpora. Here I will do this on the class (or rather *type*) level, but one could do a similar analysis also on the lexical level. Unfortunately, as explained in the previous section, in the context of this thesis the anlysis can only remain limited to a comparison of nominal type distributions. The adjectival and verbal types that are used in the qualia structure derivation cannot be trusted to give a proper indication of lexical semantic structure.

Nominal types are used in various (thematic) roles, of which the noun-noun modification ones (**nom_mod** and **nom_mod_of**) are the most frequent ones. I will therefore concentrate on the analysis of these two roles. The task of the analysis is to find among the number of semantic types that fill these roles, one or more types that are 'significant' enough to be judged as a *default* type for that role, given the particular class of lexical items under consideration. *Significant* then means a combination of both high frequency and a high MI score. Preferably we would like the default type to remain stable between different corpora, but given the diversity of domains and sublanguages this in fact is not likely to happen. Instead we will find 'defaults' for particular corpora, that can transcend into other corpora that are close enough in domain and/or language use.

`grs`

Consider for instance the types occurring with type '`grs`' (**group group_social state**) in the pattern **nom_mod_of**. Select instances for this type are as follows:

grp grs	cortege faithful fleet hierarchy
grs	academia authorities department senate swiss vanguard womankind
grs sta	agency autarchy caste coalition freemasonry membership womanhood

Figure 5.10: Polysemous classes and instances for the type: `grs`

Figures 5.11 and 5.12 give the distribution of the top 10 types occurring with these nouns in pattern *nom_mod_of* (Figure 5.11 sorted on frequency, Figure 5.12 on MI). Each type is followed by a list of corresponding basic types (for further explanation of these, see Chapter 4).

type **grs** (grp grs sta)	WSJ		BROWN		PDGF	
	freq	MI	freq	MI	freq	MI
psa (act art pos)	28	4.00	14	4.20	–	–
hum (grp grs hum nat)	13	2.62	2	2.55	–	–
lap (art com loc psy sta)	–	–	8	3.38	–	–
gas (act com evt grs psy sta)	–	–	6	2.20	–	–
gsl (grs log)	4	1.74	–	–	–	–
arg (art grp)	3	1.85	–	–	–	–
arp (art psy sta)	3	0.15	–	–	–	–
gsp (grs psy sta)	3	0.78	–	–	–	–
pya (act evt psy rel sta)	3	3.48	–	–	–	–
cea (act atr com evt psy sta)	–	–	–	–	1	9.21

Figure 5.11: Top 10 types occurring with type `grs` in pattern: **nom_mod_of** (sorted on frequency)

type **grs** (grp grs sta)	WSJ		BROWN		PDGF	
	freq	MI	freq	MI	freq	MI
cea (act atr com evt psy sta)	–	–	–	–	1	9.21
reg (grp rel)	1	6.17	–	–	–	–
arg (art grp)	–	–	1	5.17	–	–
atr (atr pro sta)	–	–	1	5.17	–	–
log (art log nat)	–	–	1	4.95	–	–
prt (art fod frm hum nat prt qui sta sub)	1	4.22	1	6.56	–	–
acr (act atr evt rel sta)	–	–	1	4.08	–	–
psa (act art pos)	28	4.00	14	4.20	–	–
hue (act evt hum sta)	1	3.97	–	–	–	–
gbs (grb grs)	–	–	1	3.92	–	–

Figure 5.12: Top 10 types occurring with type `grs` in pattern: **nom_mod_of** (sorted on MI)

As we can tell from these tables, only type '`psa`' (**act artifact possession**) is a real candidate for default type. It has both a high frequency and a high MI score[4]. Instances for this type are:

`act art pos`	cartwheel means overhead portage store stowage yield
`art pos`	bullion cache cellarage chantry counterbalance dockage footage gate hacienda homestead impost incumbrance jewel keep manor net pipage prize soap vat wharfage

Figure 5.13: Polysemous classes and instances for the type: `psa`

[4] In fact, in both WSJ and BROWN it is the same noun-noun pattern (**department:N:grs store:N:psa**), which is responsible for both high frequency and high MI scores.

Observe that apart from '`psa`' the top distributions sorted by frequency and sorted by MI are completely different (for WSJ and BROWN). Also note that this default type is comparably strong in both WSJ and BROWN, it is stronger in frequency in WSJ, but stronger in MI in BROWN. Because MI is a score that is relative to corpus size and because WSJ is bigger than BROWN, this suggests that '`psa`' is actually a stronger candidate for a default type in BROWN than in WSJ. Finally, observe the completely different distribution for PDGF, which is not surprising given the completely different domain of this corpus.

`ate`

Another interesting example is the distribution for type '`ate`' (**act attribute event psychological_feature state**) in the pattern **nom_mod_of**. Select instances for this type are given in Figure 5.14. Figures 5.15 and 5.16 give a number of potential default type candidates for different corpora.

`act atr`	abstinence adherence affectation wrong wrongdoing
`act atr evt`	bending ceremony dispersion divergence formality raising
`act atr evt sta`	complication separation
`act atr sta`	acquiescence animation behavior welfare wickedness
`atr evt`	amplitude attenuation glister glitter subsequence
`atr evt psy`	noise pity

Figure 5.14: Polysemous classes and instances for the type: `ate`

type **ate** (act atr evt psy sta)	WSJ		BROWN		PDGF	
	freq	MI	freq	MI	freq	MI
hue (act evt hum sta)	411	5.45	11	3.40	–	–
coa (act com evt pro rel psy sta)	–	–	–	–	98	4.46
pos (pos pro psy sta)	11	4.46	–	–	–	–
pro (pro)	–	–	–	–	10	3.43
gsl (grs log)	–	–	9	2.19	–	–
sub (sub)	5	2.78	–	–	–	–
nat (nat sub)	–	–	3	3.44	–	–
art (art sta)	2	3.97	–	–	–	–
cae (act art com psy sta)	2	0.58	19	3.64	–	–
gsa (act art atr com grp grs sub)	–	–	2	3.84	–	–

Figure 5.15: Top 10 types occurring with type `ate` in pattern: **nom_mod_of** (sorted on frequency)

type **ate** (act atr evt psy sta)	WSJ		BROWN		PDGF	
	freq	MI	freq	MI	freq	MI
saa (art atr sub)	1	6.75	–	–	–	–
sta (sta)	–	–	–	–	1	6.06
hum (grp grs hum nat)	1	6.05	1	3.07	–	–
hue (act evt hum sta)	411	5.45	11	3.40	–	–
avt (act art atr evt pho pro sta)	1	2.70	1	5.35	–	–
coa (act com evt pro rel psy sta)	–	–	–	–	98	4.46
pos (pos pro psy sta)	11	4.46	–	–	–	–
act (act)	1	4.26	–	–	–	–
art (art sta)	2	3.97	–	–	–	–
gsa (act art atr com grp grs sub)	–	–	2	3.84	–	–

Figure 5.16: Top 10 types occurring with type `ate` in pattern: **nom_mod_of** (sorted on MI)

First, the type `'hue'` (**act event human state**) is very frequent and also strong in MI for WSJ, with some significance also for BROWN. Select instances for this type are:

`act evt hum sta`	failure success
`act hum`	anathema apotheosis arrival bacchanal best catcher wingback
`evt hum`	breaker casualty producer quack runaway squirt stunner transient
`hum sta`	ascendant ascendent celebrity felon handmaiden hangover self servant

Figure 5.17: Polysemous classes and instances for the type: `hue`

Further, we also find four more default type candidates for separate domains. Both types `'pos'` (**possession process psychological_feature state**) and `'art'` (**artifact state**) are strong for WSJ, type `'gsa'` (**act artifact attribute communication group group_social substance**) is strong for BROWN and finally, type `'coa'` (**act communication event process psychological_feature relation state**) is strong for PDGF.

Obviously, the analysis just presented depends on many factors, with lots of variables, like accuracy of stemming, part-of-speech tagging, pattern matching, functional analysis etc. Therefore, we can take such an analysis only as an approximation. Nevertheless, it seems that given a large enough corpus, these variables are somewhat smoothed out and a (somewhat fuzzy) image appears of lexical semantic structure as it develops in real language.

Chapter 6

Underspecified Semantic Tagging

6.1 Introduction

Semantic tagging for the most part has focused on the disambiguation of homonyms, which lends itself very well to an approach that is rather similar to part-of-speech tagging: given n lexical items each with m senses apply linguistic heuristics and/or statistical measures to pick the most likely sense for each lexical item (see eg: [Yarowsky, 1992] [Wilks and Stevenson, 1996]). Such an approach, however, blurs the distinction between 'related' (*systematic polysemy*) and 'unrelated' senses (*homonymy*). Although homonyms need to be tagged with a disambiguated sense, this is not necessarily so in the case of systematic polysemy.

Lexical items that have a number of systematically related interpretations, in many cases carry those interpretations simultaneously, referring to complex semantic objects that may be in different conceptual domains at the same time. Consequently also, those interpretations are not always discretely identifiable and therefore not enumerable as a set of simple senses. These are the theoretical foundations to an *underspecified* approach to semantic tagging, which would perform a context dependent lexical semantic interpretation, while preserving information on systematic polysemy for further (discourse) processing. This chapter discusses these ideas further and presents some experiments for implementing them, using probabilistic classification on the underspecified, lexical semantic representations (qualia structures) that were described in Chapters 4 and 5.

6.2 Two Problems in Semantic Tagging

6.2.1 Multiple Reference

On the level of discourse analysis we try to distill from a text the semantic objects that are being referred to and how they interrelate (see for instance [Kamp and Reyle, 1993]). We say that semantic objects are introduced in the discourse as *discourse referents* and can be consequently referred to by *discourse markers* like pronouns and definite NP's. Consider the following simple example:

(37) [A man]$_{NP}$ walked in. He smiled.

Informally, we take the first sentence to introduce a discourse referent that corresponds to the NP 'a man' and which is picked up by the discourse marker 'he' in the second sentence. Now consider the following example from the BROWN corpus:

(38) [**A** **long** **book** **heavily weighted** **with** **military technicalities**]$_{NP}$, in this edition it is neither so long nor so technical as it was originally.

The two discourse markers 'it' refer back to an NP that expresses more than one interpretation at the same time[1]. The head of the NP (book) has a number of systematically related senses that are being expressed simultaneously: the informational content of the book (**military technicalities**), its physical appearance (**heavily weighted**) and the events that are involved in its construction and use (**long**).

If we assume that in a given NLP system discourse analysis occurs after semantic tagging this example would pose a potential problem. Discourse analysis should have access to all systematically related interpretations, whereas a traditional semantic tagger would only allow one discrete sense. Instead, I propose a form of semantic tagging that will preserve polysemous knowledge through underspecification of the lexical semantic knowledge involved.

Other researchers have pointed out the same problem. For instance, [Charniak, 1993] (quoting [Schütze, 1992]) gives the following example from *The New York Times*:

[1] Compare recent, corpus-based research on definite descriptions [Poesio and Vieira, 1998], in which it is observed that in many cases human annotators cannot decide between interpretations for certain discourse referents.

(39) In Texas, Williams, a millionaire businessman with interests in oil, cattle and banking, was hurt among voters who considered ethics an important issue.

As Charniak points out:

> "The question is, does 'interests' refer to *economic share* or *special attention*? Perhaps it refers to both. ... The natural solution to this problem is to put such words in more than one class. But this is remarkably close to the classical assumptions, except that the number of classes to which a word belongs replaces the number of senses it is deemed to have."

This remark, in fact, goes to the heart of the thesis and the idea of underspecified semantic tagging. Putting polysemous nouns in more than one class is indeed only a superficial solution that does not change anything fundamentally. Instead, and that is the approach advocated throughout this thesis, the representations (or type definitions) of the classes themselves need to be organized differently. While traditional ontologies use simple type definitions that correspond one on one to individual classes, CoreLex makes use of complex (underspecified) type definitions that involve more than one class. As explained in Section 4.3.2, underspecified semantic types are organized in such a way that the use of one class automatically involves the use of others.

6.2.2 Sense Enumeration

An intimately related problem, also pointed out by Charniak, is the identification of discrete word senses. Example (39) involves one occurrence of a lexical item ('interests'), for which it is unclear if there are really one or two senses at hand. In fact, we decided that instead there may be one complex one that corresponds somehow to two different, however simultaneous, interpretations. A similar problem exists also in deciding about a possible number of senses between different occurrences of lexical items. Charniak gives the example of the verb 'to open':

> "Does 'open an eye' have the same meaning of 'open' as 'open a box'? What about 'open one's mouth'? Is there any reason to make a sharp distinction between them?"

A number of researchers have answered this question negatively (see eg: [Pustejovsky, 1995] [Kilgarriff, 1992]). Consider these examples from BROWN:

(40a)	**fast**	run-up (of the stock)
(40b)	**fast**	action (by the city government)
(40c)	**fast**	footwork (by Washington)
(40d)	**fast**	weight gaining
(40e)	**fast**	condition (of the track)
(40f)	**fast**	response time
(40g)	**fast**	people
(40h)	**fast**	ball

Figure 6.1: Some occurrences of 'fast' in BROWN

Each use of the adjective 'fast' in these examples has a slightly different interpretation that could be captured in a number of senses, reflecting the different syntactic and semantic patterns. For instance:

'a fast action'	((40a), (40b), (40c), (40d))
'a fast state of affairs'	((40e), (40f))
'a fast object'	((40g), (40h))

Figure 6.2: A set of distinct senses for 'fast'

On the other hand, all of the interpretations have something in common also, namely the idea of 'speed'. It seems therefore useful to *underspecify* the lexical meaning of 'fast' to a representation that captures this primary semantic aspect and gives a general structure for its combination with other lexical items, both locally (in compositional semantics) and globally (in discourse structure).

6.3 Coercion and Semantic Tagging

Both the *multiple reference* and the *sense enumeration* problem show that lexical items mostly have an indefinite number of related but highly discourse dependent interpretations, between which we not always need to disambiguate. Accordingly, semantic tagging should be a first step in the interpretation process by assigning each lexical item an underspecified representation of all of its systematically related 'senses'. Semantic tags are therefore more like pointers to the probabilistic knowledge representations that were presented in Chapter 5, which can be seen as underspecified lexical meanings.

Then, if some form of disambiguation is called for, lexical items are tagged with sufficient lexical semantic knowledge to resolve all kinds of coercions, or shifts in meaning, as discussed in previous chapters. These shifts in meaning can occur: 1. between dotted elements (basic types) of the underspecified type; 2. between dotted elements and semantic knowledge that is buried deeper in the qualia structure. The semantic tagger supplies this information, while discourse based inferences, implemented for instance through abduction, derive a context dependent interpretation.

Shifts in meaning of the first kind are somewhat similar to traditional sense disambiguations for homonyms. However, with polysemous nouns we need to preserve not just the interpretation of the 'local' discourse, but also all related interpretations that may be needed for further discourse processing. As we have seen before, it may very well be the case that one aspect of the noun is referred to in a particular piece of the discourse, while another related aspect is referred to further on. I repeat example (41) for illustration:

(41) He threw out the book after reading it.

book: **information$\bullet$physical**

Now compare the somewhat similar example (42) with 'rabbit', a polysemous noun represented with an *open* dotted type (combining basic types that are unlikely to occur simultaneously in discourse).

(42) John likes to eat rabbit. ... He shot it in the hills nearby.

rabbit: **animal$\circ$food**

In both examples, discourse analysis would need to have access to all aspects of the polysemous noun, but in each case the analysis itself will be slightly different. Whereas the analysis in (41) will be straightforward, in the case of (42) it will involve one extra step of relating the **animal** aspect to the (prepared) **food** aspect. A step, in fact, that involves deeper semantic knowledge represented in the qualia structure, in the nature of verbs that express the 'prepare' relation between animals and the food you can make out of them.

This, of course, is actually an example of the second kind of meaning shift, illustrated also by the following examples (repeated from Chapters 2 and 3):

(43a) BMW announced a new model.

(43b) The Boston office called.

In each case, one ore more extra inference steps are needed that involve qualia structure information. In (43a), that 'BMW is a company that employs people who can make announcements'; in (43b), that 'an office is a place where people work who can make phone calls'.

The two kinds of shift in meaning are very different in nature, but nevertheless intimately related. The first kind (shifts between basic types in the dotted type) is theory driven and therefore completely predictable. The second kind (shifts involving the rest of qualia structure) is data driven and largely unpredictable. The two are connected, however, through representation over time. Given enough data, shifts of the second kind can stabilize and may warrant extension of the dotted type with one or more extra basic types, that can be obtained from qualia structure shifts. For instance, in the case of example (43a), we may conclude, over time, given enough similar examples, that (proper) nouns like 'BMW' can be described by a dotted type that includes the basic type **human**. (Something, which, in fact, we did: See Section 4.3.2.)

The extension of a dotted type with newly acquired additional basic types really is a shift in meaning that is truely dynamic, reflecting a change over time in the outlay of the ontology itself. To monitor shifts like these we can utilize semantic tagging by comparing, on the one hand, shifts between basic types (in the dotted type) and other shifts arising from the rest of qualia structure. In the next sections I present some experiments in actually implementing this by use of probabililistic classification.

6.4 Classification for Semantic Tagging

Using pattern matching for the acquisition of lexical semantic structure, as discussed in chapter 5, we can implement an experimental underspecified semantic tagger that employs the CoreLex ontology of systematic polysemous classes and corresponding underspecified semantic types. This ontology was built on a dataset of around 40,000 nouns, which obviously does not cover every noun in the English language. New words are created every day, while existing words acquire new interpretations. To handle this constant change and expansion of language one needs to design algorithms that classify words into existing or emerging classes. Here, I describe one such algorithm that classifies lexical items strictly according to their *linguistic context*, that is, according to their distributions within certain linguistic constructions.

Knowledge acquisition methods like classification can range from knowledge poor methods, assuming few or none *a priori* knowledge sources, to knowledge rich ones. Purely statistical, connectionist research in lexical knowledge acquisition is knowledge poor (for instance [Brown *et al.*, 1992] [Schütze, 1992] [Schütze, 1994]), whereas research that relies more on linguistic structure tends to be richer in knowledge (for instance [Hindle, 1990] [Brent, 1993] [Pustejovsky *et al.*, 1993] [Pereira *et al.*, 1993]). The approach I take here assumes the following sources of a priori knowledge: *part of speech tags; morphological stems*; some *syntactic structure*; sets of *semantic classes* and *semantic roles*.

As pointed out in [Schütze, 1994], such a priori knowledge sources are symbolic (discrete, knowledge based) representations of linguistic *competence*, while subsymbolic (continuous, statistical) representations give an account of *performance*. Schütze then goes on to argue that for the study of *performance* subsymbolic representations are superior. Although I subscribe to this opinion, it needs to be observed also that for a study of *performance* to contribute to our understanding of *competence* we first need such a theory to start with and then see how empirical data verify this theory or not. This implies that, in fact, we need both a knowledge based (theory driven) and an empirical (data driven) study of lexical semantics, both to establish and to verify the theory. It is such an approach to semantic classification that is taken in this chapter.

Also, the use of a priori knowledge sources can change something fundamental in the basic task of the classification algorithm, that is: grouping together similar objects. If the algorithm relies on a predefined set of classes of objects, this task is rather different than without such a set. In the first case we are really talking about *classification*, whereas in the second case we are rather referring to the task of *clustering*. The approach taken here specifically assumes a set of predefined semantic classes as defined in the CoreLex ontology. In this I follow previous work by a number of researchers in semantic classification, word sense tagging and disambiguation [Yarowsky, 1992], [Resnik, 1992], [Grefenstette, 1996], who point out the need for an a priori classification scheme, like the ROGET thesaurus, as a *gold standard* to compare classification results with.

6.4.1 Classification Algorithm

For the purposes of this thesis, semantic classification is based on similarities in linguistic context between nouns (or lexical items in general). This approach follows directly that of [Hindle, 1990], but includes a far richer set of linguistically motivated attributes (syntactic constructions and functional relations) as discussed in Chapter 5. The linguistic knowledge that was described there is used here by a simple classification algorithm, as given in Figure 6.4, that compares the linguistic contexts of nouns and *classes* of nouns.

At the core of the algorithm is the computation of a similarity score between each noun and each (CoreLex) class. For this purpose I use the *Jaccard measure* that compares objects relative to the attributes they share [Grefenstette, 1994]. Here, attributes are the different (thematic) roles of a noun: **agent, theme, nom_mod, adj_mod,** etc. The Jaccard measure is defined as follows:

$$\frac{A}{A + B + C}$$

A	:	attributes shared
B	:	attributes unique to object 1
C	:	attributes unique to object 2

Figure 6.3: Jaccard measure

The Jaccard scores for all CoreLex classes are sorted and the class with the highest score is assigned to the noun. If the highest score is equal to 0, no class is assigned. Time complexity of the algorithm is *linear*, because the number of classes is constant.

```
foreach noun_stem
    foreach class
        compute Jaccard score using class data and noun_stem data;
    end
    output class with highest Jaccard score, unless highest = 0;
end
```

Figure 6.4: Classification Algorithm

Jaccard Score: A Simple Example

Consider a simple example from the STOCKS text. The type (or class) of the noun 'result' is unknown (it does not occur in CORELEX), so we need to compute a similarity between this noun and all CORELEX classes by comparing their 'linguistic context data', using the Jaccard measure. There are two classes that have such data (attributes) in common with the noun 'result': 'pos' and 'tme'. The attributes for these two classes are given in Figure 6.5, while those for 'result' are given in Figure 6.6. Each row gives a noun (or class) with its total frequency in the given text, the attributes (syntactic constructions and functional relations) in which it appears and the values for those attributes. Each value is further represented through its part of speech (*pos*), the frequency of co-occurrence in the text of this noun/class and this attribute_value pair and a mutual information (*MI*) score for the pair.

class	*class-freq*	*attribute*	*value data*			
			value	*pos*	*freq*	*MI*
pos	3	adj_mod	quarterly	adj	1	5.64
pos	3	**adj_mod**	**second**	adj	1	4.54
pos	3	theme	report	verb	1	4.54
tme	5	adj_mod	heavy	adj	1	4.44
tme	5	**adj_mod**	**second**	adj	1	4.03
tme	5	nom_mod	trading	noun	1	4.03
tme	5	theme	look	verb	1	4.03
tme	5	time	profit	noun	1	5.13

Figure 6.5: Data on classes 'pos' and 'tme' in STOCKS

noun	*noun-freq*	*attribute*	*value data*			
			value	*pos*	*freq*	*MI*
result	3	adj_mod	disappointing	adj	1	5.64
result	3	**adj_mod**	**second**	adj	1	4.54
result	3	nom_mod	quarter	noun	1	4.54

Figure 6.6: Data on noun 'result' in STOCKS

The noun 'result' has only one attribute in common with each of the two classes: **adj_mod** with the value 'second', while a number of other attributes are unique to each. We compute the Jaccard score between the noun and the two classes on the frequencies of their attributes (the column marked *freq*). Given the formula in Figure 6.3, this gives a value of 2 for A (the number of attributes shared * their frequencies), 2 for B (number of attributes unique to 'result' * their frequencies), 2 for C in the case of 'pos' (number of attributes unique to 'pos' * their frequencies) and 4 in the case of 'pos' (number of attributes unique to 'tme' * their frequencies). The resulting computation gives the following values: 'pos' (0.33) and 'tme' (0.25), assigning 'result' the class 'pos'.

'pos' : A = 2; B = 2; C = 2	Jaccard = 2/6 (0.33)
'tme' : A = 2; B = 2; C = 4	Jaccard = 2/8 (0.25)

Figure 6.7: Jaccard scores on 'result' with classes 'pos' and 'tme' in STOCKS

6.4.2 Evaluation

The classification process is evaluated in terms of precision and recall figures that are computed on the classification of those nouns that are in CoreLex, because their class can be checked automatically. For the purposes of evaluation each corpus (*text*) is classified 'on itself', meaning that class data from the text itself is used to classify individual noun stems. In order to explain what this exactly means and also to illustrate some of the implications of such an approach consider the following example.

Let us assume that in ideal circumstances, for a given text there are 100 nouns and 100 classes to which these nouns belong (according to the CoreLex database). Each noun belongs to one class and each class has only one member. Also, there is only one attribute_value pair for each noun and (consequently) each class. Under such conditions classifying each noun in the text by using data from the text itself is straightforward. Comparing each noun to each class will yield a result in which all nouns will be most similar (in fact identical) to the class to which it belongs (according to the CoreLex database). This is what I call a 'reconstruction' of the CoreLex classification. In ideal circumstances as just described such a reconstruction can be performed with a 100% accuracy on both precision and recall. In practice, however, things will be less ideal and precision and recall on the reconstruction will be much lower.

For instance, not every noun in a text will be in the CoreLex database to start with and its class will be therefore unknown. Also, most classes will have more than one member and both nouns and classes will have more than one attribute_value pair. Figure 6.8 gives an overview of the precision and recall scores for a number of corpora. Results vary depending on size (number of potentially different attribute_value pairs) and homogeneity (number of potentially overlapping attribute_value pairs).

The first and second columns give an indication of size of the corpora, in terms of number of tokens and different noun stems. The third column shows what percentage of noun stems for this particular corpus can be found in CoreLex, while the fourth column gives the percentage of those noun stems that are in CoreLex and can actually be classified, because linguistic data for them is available (this depends on the accuracy of the pattern matcher as discussed in Chapter 5). It is the classification of these noun stems that has to be reconstructed, of which column five gives a measure of precision.

corpus	*tokens*	*nouns*	*in* CORELEX	*in* CORELEX*, classified*	
		stems	recall	recall	precision
STOCKS	846	85	47.06	36.47	90.32
ORWELL	6190	577	63.60	32.93	76.32
MR	23059	493	55.58	44.62	63.64
DARWIN	168620	2370	58.06	40.08	45.89
PDGF_SMALL	137079	2310	42.79	35.56	42.95
WSJ_SMALL	146487	4033	67.89	52.17	46.06
BROWN_SMALL	146288	4888	67.31	49.22	45.76

Figure 6.8: Precision and recall evaluation

Precision works out much better for smaller texts than for larger ones, because of increasing homogeneity of smaller texts. Homogeneity plays an important role as may be obvious from the numbers for the DARWIN, PDGF_SMALL, WSJ_SMALL and BROWN_SMALL[2] corpora, all of which are comparable in size. The corpus with the lowest precision (PDGF_SMALL) is indeed the least homogeneous of these four[3].

6.4.3 Parameters on Classification

We can increase the precision of the classification algorithm significantly by taking into account a number of parameters that will smooth out the effects of low frequency. First, we can classify on *mutual information* scores instead of raw frequencies. Secondly, we can take into account *relative class size*, reflecting the fact that some classes occur more frequently than others. This is specifically interesting in respect to corpora that correspond to very specific domains. Additionally, we can investigate the effects of classifying over *classes* of attribute_value pairs instead of these pairs themselves and of putting weights on different sources of *linguistic information* (NP, VP, AP). Of these four parameters, only the first two are effective in increasing precision, the other two rather reduce precision.

Increase in precision is important in the context of increasing the accuracy of the tagging system, at the core of which is the classification algorithm. So, in order to yield better results on identifying the nouns and interpretations that are illustrations of the topic of this thesis (systematic polysemy and underspecification) it is worthwhile from an empirical standpoint to investigate the different parameters. From a theoretical point of view, however, things are much less clear. For instance, it is not completely obvious what the theoretical contribution of mutual information is to the identification of systematic polysemy and underspecification. There is, in principle, no difference in the computation of mutual information for a systematic polysemous and a non-systematic polysemous (homonymous) noun. However, in theory, a systematic polysemous noun will have a range of similar mutual infor-

[2]PDGF_SMALL, WSJ_SMALL and BROWN_SMALL are 'small' versions of the larger corpora of the same names that were obtained by taking the first 6,000 or 7,000 lines of these.

[3]Sophisticated methods for deriving measures of homogeneity of corpora are discussed in [Kilgarriff, 1997b]. Future research should apply these methods in order to analyze results more accurately.

mation scores for its co-occurrence with different lexical items. A homonymous noun on the other hand will have mutual information scores that are widely different for each of its homonymic senses.

Figure 6.9 shows performance of the classification with three parameters: mutual information (*MI*), relative class size (*size*) and classes of attribute_value pairs (*classes*).

corpus	*in* CORELEX	*in* CORELEX*, classified*				
	recall	recall	precision	precision	precision	precision
			(freq)	(MI)	(size)	(classes)
STOCKS	47.06	36.47	90.32	90.32	90.32	90.32
ORWELL	63.60	32.93	76.32	77.37	81.05	81.05
MR	55.58	44.62	63.64	66.36	72.73	64.55
DARWIN	58.06	40.08	45.89	53.68	68.53	65.79
PDGF_SMALL	42.79	35.56	42.95	52.66	67.42	61.30
WSJ_SMALL	67.89	52.17	46.06	51.00	62.60	51.62
BROWN_SMALL	67.31	49.22	45.76	50.29	61.76	50.58

Figure 6.9: Precision using several parameters

Mutual Information

Mutual information was explained already in Chapter 5. Again, in summary, this score shows how strongly connected two (classes of) lexical items are, relative to the rest of the corpus. Using mutual information scores gives an increase on precision between 1% and almost 10% (PDGF_SMALL).

Relative Class Size

Another effective idea for optimizing performance is to take into account *relative class size*. Not all classes occur as many times as others in a text. Frequently occurring classes are for instance **artifact** and **human**. These are also classes with most instances in CoreLex. Also, classes that are more domain dependent (for instance **cell** and **group_biological**) occur more frequently in some texts than in others. By applying a measure of size to each class, based completely on frequency, we can make some classes more prominent than others, which gives a much higher precision in classification. The increase for larger corpora is between 10% and 15%. Also note that the highest increase is obtained for the most domain specific corpora (DARWIN and PDGF_SMALL), which confirms our intuitions.

Classes

It may also seem to be a good idea to classify over classes of attribute_value pairs instead of over those pairs themselves. This groups together co-occurrences that are similar from a semantic point of view and would otherwise be split up. However, as the numbers in Figure 6.9 show, precision actually goes down significantly. A reason for this could very well be the lack of coherence within the classes that correspond to the CoreLex types. Because of their, although related, but still different interpretations, a number of varying contexts will be accumulated, resulting in classification attributes that are not 'focused' enough for the classifier to perform well.

Linguistic Sources

Finally, we look at the effect of *source* of linguistic information. Figure 6.10 compares results for the four larger corpora with weights applied to information from within AP's (adjectival modification), VP's (thematic roles) and NP's (nominal modification and thematic roles with nouns). We find that all corpora have similar behaviors, with decreasing results almost overall. There is only a slight increase on NP information for three corpora (all except DARWIN), with a weight between 2 and 5.

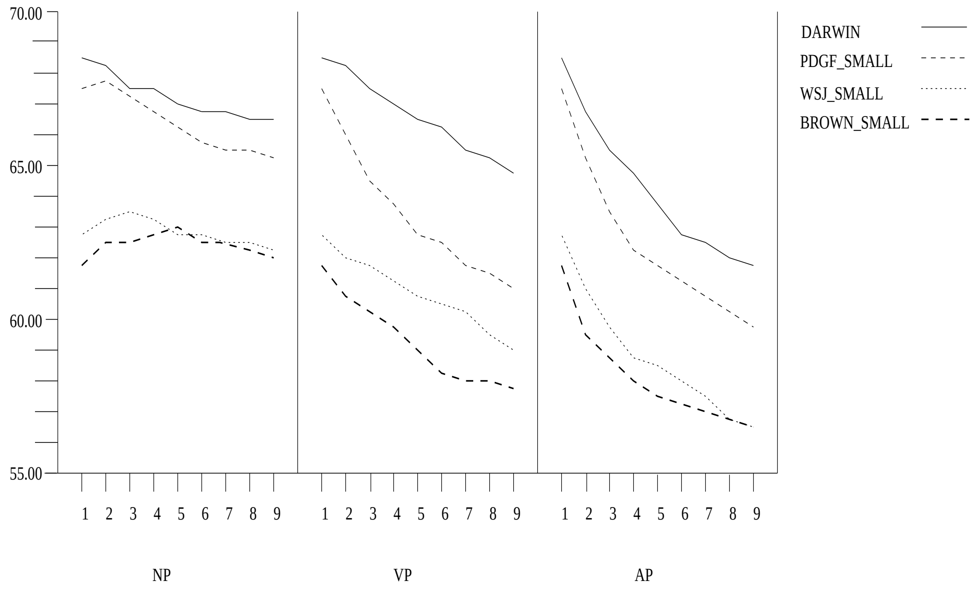

Figure 6.10: Precision with weights on *linguistic sources*

6.5 Experiments in Underspecified Semantic Tagging

In the previous section I presented a quantitative evaluation of the probabilistic classifier that is at the heart of the semantic tagger discussed in this chapter. Here, on the other hand, I will give some qualitative results. This includes the acquisition of new instances of CoreLex types, specification between basic types that constitute each underspecified CoreLex type, and shifts in the meaning of polysemous nouns that go beyond their CoreLex type.

Identifying Instances

Classification for tagging unknown nouns (those nouns in a text that are not in CoreLex) is, in fact, a very general task. Unknown nouns need to be handled by any (semantic) tagger, independent of the type of tags, sense based or underspecified. As discussed above, however, for this to work well, classes need to be somehow coherent, which obviously is a problem in dealing with polysemous nouns. Nevertheless, by showing that new instances for classes can indeed be found gives some proof that the classes are coherent at least in respect to their polysemy. That is, new instances can be found that are polysemous in similar ways.

Identifying Interpretations

Classification is further used for the identification of coercions or shifts in meaning as discussed in Section 6.3. Coercions occur either within the dotted type, or from within the qualia structure. The first application is to tag a noun with a basic type that is part of its dotted type. This is the contextual determination of a polysemous interpretation, that leaves relations with other interpretations open to further discourse processing. The second application is to tag a noun with a basic type that is not part of its dotted type. This identifies a meaning shift to a novel interpretation that originates in qualia structure. The novel usage of a noun is observed by a collocation that violates current selection restrictions. This involves the defaults that were discussed in Section 5.4.

6.5.1 Some Results on Classifying Unknown Nouns

In all experiments described in this, and also in the next section, classification is performed on texts by comparing them with data from large corpora, to be specific: BROWN and WSJ. As mentioned before, these corpora represent certain domains that are partly overlapping, but are also for a large part mutually exclusive.

STOCKS

For instance, compare in Figure 6.11 the results of classifying unknown nouns in STOCKS with both BROWN and WSJ data. In the case of BROWN, 3 and in the case of WSJ, 2 unknown nouns were classified in a probabilistic fashion. Of these, however, none are shared between the results from both corpora. Also, both classifications are dissimilar in rate of accuracy, but this is something that is hard to put in exact numbers. At most, we can pin down quite accurately the clear misses: 'morning', classified as being of type 'hum' (**group group_social human natural_object**). The hits are as follows: 'cap' is short for 'capitalization', which is indeed an **act** or **event**; 'value' and 'index' are correctly classified as similar to 'bargain' and 'price'; 'investment' is judged correctly to be (somewhat: 0.10) similar to 'research' and 'technology'.

class (with instances from STOCKS)	*derived*	
	BROWN	WSJ
aev : act, event		0.25 cap
(crash)		
hum : group group_social human natural_object	0.01 morning	
(analyst; consumer; investor; participant; strategist)		
pom : act artifact attribute communication possession	0.60 value	
(bargain; price)	0.29 index	
pya : act event psychological_feature relation state		0.10 investment
(research; technology)		

Figure 6.11: Probabilistically derived class instances in STOCKS

NREN

In the following, we look at texts that are either somewhat similar to both domains, or rather dissimilar. First, we look at NREN, a U.S. government document on the developement of national information networks, which reflects the domain of information technology that is, into some respect, part of both BROWN and WSJ. However, here we have to take into account also the historical developement of this field, which is still quite recent and is consequently not reflected very extensively in the somewhat dated BROWN corpus. Its representation is most likely quite better in the more recent WSJ corpus. Nevertheless, the results are largely comparable, although again strikingly different, as shown in Figure 6.12.

class (with instances from NREN)	*derived*	
	BROWN	WSJ
`atp` : attribute psychological_feature state		0.50 security
(capability; purpose)		0.44 characteristic
`cea` : act attribute communication event psychological_feature state	0.09 security	
(attention; question; restriction)		
`ent` : entity	0.58 carrier	0.99 sector
(entity)	0.54 trunk	
`gsl` : group_social location_geographical	0.28 node	
(nation)		
`pya` : act event psychological_feature relation state		0.06 education
(project; research; technology)		
`rel` : attribute human psychological_feature relation state	0.18 response	
(aspect)		

Figure 6.12: Probabilistically derived class instances in NREN

For BROWN, 5 and for WSJ, 4 unknown nouns were classified. Of these, only 1 ('security') is shared, but it is classified differently between both domains. There are no clear misses for both BROWN and WSJ, although the classification of 'security' is questionable. Both the classification in BROWN and in WSJ can be argued for. In fact, the two classes that 'security' is classified under (`'atp'` and `'cea'`) have a strong relationship to each other. The basic types that correspond to the class `'atp'` are included in those that correspond to the class `'cea'`. This example again shows that borderlines between classes are not clearcut.

In the case of NREN we also see an instance ('entity') of a very general class (**entity**). For BROWN, there are two nouns ('carrier' and 'trunk') that are identified as similar to this class. Although this is not a clear miss, it is hard to tell if it is a real hit either, exactly because the **entity** class is so general. However, for WSJ, the noun 'sector' is judged highly similar to this class, which tells us something about the more specific meaning of 'entity' in the NREN text. Here, namely, the (normally very general) noun 'entity' actually means something like 'sector' or 'part' of a network, which is the general topic of the NREN text. This example thus nicely demonstrates the intimate connection between the identification of new instances for known classes and of finding new interpretations for the instances of these classes.

MR

Next we look at a text that is largely different from both BROWN and WSJ. Like NREN, the MR text is on information technology, but instead of taking the form of a scientific or journalistic report it is a very technical user manual with a clear focus on the 'Apple Macintosh' computer. This is a subject that is too specific to be reflected in either domain and consequently the results on both are quite bad. In Figure 6.13 we therefore compare these results (for selected classes only) with a classification on MR itself, which works out significantly better, mainly because of the homogeneity of this text.

For instance, consider the two CoreLex classes `'avr'` and `'evt'`. There are no derived instances at all for these classes, if we use BROWN or WSJ data, while results using MR data are excellent. Similarly, satisfying results for MR and dissatisfying ones for BROWN and WSJ are obtained for most other classes.

class (with instances from MR)	*derived*		
	MR	BROWN	WSJ
acp	0.97 resource	0.16 clicking	0.34 field
(activity; operation; search; use)	0.49 support	0.17 glass	0.12 shape
	0.18 system	0.14 field	
act	0.95 exchange	0.03 bottom	
(cleaning; expansion; insertion; move)	0.95 medium	0.02 opening	
	0.95 repairing		
	0.22 stage		
aev	1.00 magnification		0.33 force
(substitution)	0.29 family		0.33 middle
	0.14 sample		0.33 test
art	0.35 ball	0.01 background	0.04 mechanism
(adapter; clock; cloth; connector; desk)	0.33 item	0.01 beginning	0.03 closeview
	0.22 bar		0.02 top
	0.18 table		
	0.18 work		
	0.16 card		
avr	0.66 adjustment		
(click; map; modification)	0.52 image		
	0.44 change		
	0.43 area		
coa	0.51 protection	0.58 time	
(capitalization; comment; communication)	0.30 initialization		
evt	1.00 model		
(beep; drip; malfunction)	1.00 restart		
	1.00 return		
grs	0.95 anyone	0.10 product	0.08 record
(hierarchy; household; people)	0.72 everyone	0.09 mark	0.06 start
	0.69 link		
	0.53 ll		
	0.34 chain		

Figure 6.13: Probabilistically derived class instances in MR

Larger Corpora

The results we have looked at in this section were derived from relatively small corpora. The number of classes and noun instances involved here are accordingly small and can be evaluated manually. If we want to evaluate results on larger corpora however, we need a different, more automatic approach, which should be goal directed. For instance, we could evaluate performance in the context of an information-extraction program, that would perform better or worse, depending on the correct identification of unknown nouns in a given domain and corpus of relevant texts. The design and implementation of such an embedded context for the classification algorithm discussed here is, however, beyond the direct scope of this thesis and is therefore left to further research.

6.5.2 Some Results on Identifying Interpretations

The same strategy for identifying new instances of classes by classifying unknown nouns according to their linguistic context can be used also to identify interpretations of nouns that are more *specified* than their underspecified semantic type. This section presents some experiments in using the classification to tag nouns in respect only to their 'local context', that is, the sentence in which they occur[4]. This strategy is used in two ways: 1. to identify the most appropriate (statistically likely) basic type out of the basic types that constitute the dotted (underspecified) type of a noun; 2. to identify the most appropriate basic type, ranging over all basic types, independent of the dotted type of a noun. Approach 1. identifies more specified interpretations within an underspecified one, whereas approach 2. identifies new interpretations, outside of the underspecified one.

The results of the experiments that are discussed in this section are indicative of the rather interesting potential possibilities of such an approach to lexical semantic analysis, but unfortunately the current level of accuracy is not sufficient for a full scale analysis. These results are therefore to be taken as merely illustrative and not necessarily representative of all of the results currently obtained.

[4] Ideally, 'local context' should be a coherent discourse segment, but this has not been implemented.

Identifying Interpretations within Dotted Types

The research reported in this section is similar to, but should not be compared directly with work on word sense disambiguation (WSD), as for instance reported in [Yarowsky, 1992] [Wilks and Stevenson, 1996] [Dagan *et al.*, 1997]. First, because most of the WSD research is concerned with homonyms, where discrete senses are more easily discerned. Results are therefore hardly comparable with the task put forward here, that is, to distinguish word sense differences on a much more fine-grained level. As observed in [Kilgarriff, 1997a] – quoting Yorick Wilks (personal communication):

> "The state of the art is such that WSD with coarse-grained senses, or homographs, is attainable at well over 90%, whereas for fine-grained WSD even human inter-annotator agreement is very low."

Secondly, whereas the research discussed here deals with a large number of nouns, spread over 126 different semantic types, in order to show the flexible and dynamic nature of word meaning, WSD research typically deals only with a small number of nouns, in order to show the speed and accuracy of a particular algorithm.

The results discussed here, rather illustrate the idea of highlighting a specific interpretation of a polysemous noun in context. Consider for instance sentence (44) from STOCKS:

(44) The world's most closely watched stock index rose in morning trading, plunged to a loss of more than 160 points in early afternoon, then roared back to trade briefly with a 50-point gain before closing modestly higher.

In STOCKS, the polysemous noun 'gain' is tagged with the underspecified semantic type `'pas'` that includes the basic types: **act attribute possession relation state**, reflecting a polysemous meaning that can fluctuate between these basic types. In the local context of sentence (44) this noun was identified by the tagger to have the interpretation **act**, if compared with data obtained from the WSJ corpus.

Another example from STOCKS is the polysemous noun 'earnings', which is tagged with the underspecified semantic type `'pos'` that includes the basic types: **possession process psychological_feature**

state. In the local context of sentence (44) (specifically within the NP 'A batch of strong quarterly earnings'), it is tagged with the interpretation **psychological_feature**, again if compared with data from WSJ.

(45) A batch of strong quarterly earnings, thought to be key to stopping Wall Street's slump, failed to bring any lasting relief.

A final example, also from STOCKS is 'price', which is tagged with 'pom', including: **possession process psychological_feature state**. In sentence (46) (specifically within the VP 'consumer prices rose'), it is tagged with the interpretation **possession**, compared to WSJ.

(46) Analysts said the market was boosted initially by bargain-hunting, and the mood on Wall Street improved slightly following news that consumer prices rose just 0.1 percent in June, soothing fears that inflation pressures were brewing and that the Federal Reserve would raise interest rates.

Identifying New Interpretations

Next to using classification for identifying context specific interpretations of polysemous nouns, given their dotted types, we can use this same strategy for identifying new interpretations that are not covered by the dotted type. Consider for instance sentence (47) from STOCKS:

(47) Stocks roared back from massive losses in the heaviest trading day ever on Wall Street as bargain hunters swooped in to buy battered shares.

In the STOCKS text the polysemous noun 'share' is tagged, in the local context of sentence (47), with the interpretation **food** if compared with data obtained from the BROWN corpus, because in BROWN 'battered' is connected most significantly (in terms of *mutual information* scores) with nouns of type **food**. Therefore, we can argue that in (47) the noun 'share' acquired the new interpretation of **food**. This by itself is of course not correct, because the adjective 'battered' is itself *homonymic* between 'beating' (as in (47)) and 'covering in a flour mixture' (as with **food**), but the example illustrates the technique of identifying potential new uses.

A better example is the interpretation of the noun 'index' in the following sentence:

(48) The world 's most closely watched stock index rose in morning trading, plunged to a loss of more than 160 points in early afternoon, then roared back to trade briefly with a 50-point gain before closing modestly higher.

In STOCKS (taking all contexts together) 'index' is tagged correctly with the underspecified semantic type 'pom' that includes the basic types: **act artifact attribute communication possession**. However, in the local context of sentence (48), this polysemous noun is tagged with a basic type (**quantity_definite**) that is different from these. Again, tagging the noun with this type is based on comparing its linguistic context with data obtained from BROWN, in which the verb 'to rise' is connected most strongly with nouns of type **quantity_definite** (filling its AGENT role).

A further example is the interpretation of 'market' in (49):

(49) In the broader market, declining issues swamped advances 1,836 to 691 on record volume of 680 million shares on the New York Stock Exchange, surpassing the old high of 652.8 million set on Dec. 15, 1995.

Here, the polysemous noun 'market' is tagged with the basic type **location_geographical**, based on its occurrence with the adjectival modifier 'broader', which in BROWN is associated strongest with nouns of such type. However, the CORELEX type for 'market' is 'gsa', which includes the basic types: **act artifact attribute communication group group_social substance**. Consequently, on the basis of usage like in (49), we may conclude that 'market' acquired the further interpretation of **location_geographical**.

We now look at some examples from a different text, DARWIN, compared with data from the BROWN corpus. This gives interesting results for the somewhat similar polysemous nouns 'catalogue' and 'list', as given in sentences (50) and (51):

(50) I will then pass on to the variability of species in a state of nature; but I shall, unfortunately, be compelled to treat this subject far too briefly, as it can be treated properly only by giving long catalogues of facts.

(51) A long list could easily be given of 'sporting plants;' by this term gardeners mean a single bud or offset, which suddenly assumes a new and sometimes very different character from that of the rest of the plant.

According to CoreLex, 'catalogue' is typed as `'cae'`, including the basic types: **act artifact communication psychological_feature state**, whereas 'list' is typed as `'com'`, including: **communication psychological_feature relation state**. However, on the basis of their usage in (50) and (51), both of these polysemous nouns are tagged as **quantity_definite**, resulting from their collocation with the adjective 'long'.

Finally, we look at an example from DARWIN that also involves a noun of type `'cae'`. If the context of 'production' in sentence (52) is compared with WSJ, it acquires the interpretation of **food**, based on the collocation of 'production' with the adjective 'domestic' on the one hand and the verb 'to raise' on the other. In fact, the context of 'production' in sentence (52) ('domestic productions are raised') gives exactly rise to the interpretation **food**.

(52) When we reflect on the vast diversity of the plants and animals which have been cultivated, and which have varied during all ages under the most different climates and treatment, I think we are driven to conclude that this greater variability is simply due to our domestic productions having been raised under conditions of life not so uniform as, and somewhat different from, those to which the parent-species have been exposed under nature.

This example, also again, illustrates the need for an underspecified view on semantic tagging, that would allow the discourse referent introduced by the NP 'our domestic productions' to be linked to the one introduced by the NP 'the vast diversity of the plants and animals which have been cultivated'. To establish coherence in discourse analysis in such a way, we obviously need to *annotate* lexical items with sufficient knowledge (collocations with default and non-default lexical items) for explaining the interpretations they acquire in context. The work explored in this thesis is intended to contribute to such an approach.

Chapter 7

Conclusions

7.1 Contributions

The main contributions of this thesis to the fields of natural language processing and computational linguistics are the CoreLex ontology and lexical semantic database, the introduction of an underspecified view on semantic tagging and the developement of finite-state methods for the acquisition of underspecified lexical semantic representations. A more detailed description of these applied contributions follows below. On a more theoretical level, the thesis made comprehensive contributions to a closer integration of studies in lexical semantics and discourse analysis by investigating the intimate connection between systematic polysemy and underspecification.

Innovative Ontology Design for Underspecified Lexical Semantic Representation

Central to this thesis is the construction and use of the CoreLex ontology and lexical semantic database. The theoretical background to CoreLex, centering around the ideas of systematic polysemy and underspecification, is an important innovation in the design of such resources. It is as far as I know, the first lexical semantic resource with a basic design that moves away from the traditional paper dictionary design of enumerations of systematically polysemous senses.

Instead, CoreLex offers a new approach, based on $\mathcal{GL}$ theory, that leaves the number and exact nature of these senses underspecified. This allows for more robust natural language processing with interpretation choices, due to systematic polysemy, that can be left open, because they need not necessarily be resolved, depending on the task involved.

Theory and Implementation of an Underspecified Approach to Semantic Tagging

The CoreLex approach to lexical semantic representation leads quite naturally to a corresponding underspecified approach in semantic tagging. As explained in different parts of the thesis, the analysis of discourse coherence, the connection between individual propositions that make up the meaning of a text, depends on the availibility of lexical semantic knowledge that can explain the connections between the propositions. For instance, this would involve expanding the meaning of discourse referents to include metonymic, metaphoric and coercive interpretations.

For this process to succeed, we need to tag, or rather *annotate*, the lexical items in the text with sufficient information that for a large part will have to be underspecified in order to accommodate multiple related interpretations, due to the systematic polysemy of the lexical items involved. This, in essence, is the implication of the strong connection between systematic polysemy in lexical semantics and underspecification in discourse that is discussed in this thesis.

Methods for Large Scale Acquisition of Underspecified Representations

The underspecified representations that are used in semantic tagging are not readily available, they need to be acquired on a large scale, connecting theory and empirical analysis. On the theoretical level, the research established the CoreLex ontology of 126 underspecified semantic types that reflects most of the examples of systematic polysemous classes of nouns mentioned in the literature plus many others that were perhaps never observed before.

On the level of empirical analysis, the thesis describes a number of different, but related efforts to connect the CoreLex ontology with various data on systematic polysemy that can be found in any corpus. Central to these is the use of probabilistic classification to find new instances of systematically polysemous classes, construct corpus-based, underspecified representations for these classes and to identify systematically related interpretations in corpora.

7.2 Future Research

The underspecified representations that are acquired from WordNet and from domain specific corpora, connect hierarchical (*paradigmatic*) and compositional (*syntagmatic*) knowledge. The acquisition processes for both of these are discussed extensively in this thesis, but given the many variables involved in each, they also leave a lot of room for improvement, which would be a main topic for future research. Additionally, the results on 'semantic shifts', that originate in the interaction between syntagmatic and paradigmatic knowledge over time, although intriguing, require much more research, in order to establish a more accurate, consistent and (preferably) automatic analysis of this process.

Large Scale Discourse Analysis for Dotted Type Settings

All of Chapter 4 is devoted to the acquisition of paradigmatic knowledge, by establishing a set of semantic types that define membership into a systematic polysemous class. The dotted type connectors that are used in these underspecified types are defined and illustrated in Chapter 3. However, in order to determine the right connector ('$\bullet$') or ('$\circ$') for all underspecified types that were established by the thesis, a large scale (corpus-based) discourse analysis, involving these types, should be carried out. This would be interesting future research, connecting the fields of lexical semantics, discourse analysis and corpus-based analysis rather strongly.

Increasing the Accuracy of Finite-State Knowledge Acquisition Methods

The acquisition of syntagmatic knowledge is discussed in Chapter 5. The analysis of modification and thematic role assignment proceeds through a cascade of finite-state techniques that include tokenizing, part-of-speech tagging, stemming, verb- and noun-group recognition and semantic typing. The resulting data is organized along the lines of a qualia structure like representation, as discussed in Chapter 3. Although the results so far give some good indications of the general possibility of acquiring lexical semantic representations in this way, in future research, the overall accuracy of the techniques mentioned should be improved.

More Accurate and Automatic Analysis of Semantic Shifts

The effect of syntagmatic knowledge on paradigmatic knowledge is discussed in Chapter 6. Differences in the (typical) composition of lexical items will, over time, result in a change of class membership for those lexical items. This is what I have called 'semantic shifts', a term with connotations in etymological studies and historical linguistics. Although this is not directly what we are after here, the term still seems appropriate, given the dynamic and hence diachronic nature of the changes in meaning from one corpus to another (especially, of course, if they are chronologically sequential) and even from one discourse (segment) to the next. Future research should determine more accurate methods for analysing these changes and find automatic ways to represent them in useful ways.

Bibliography

[Apresjan, 1973] J. Apresjan. Regular Polysemy. *Linguistics*, (142), 1973.

[Asher and Morreau, 1990] Nicholas Asher and Michael Morreau. Commonsense Entailment: A Modal Theory of Nonmonotonic Reasoning. In Jan van Eijck (ed.), *Proceedings of JELIA90; Lecture Notes in AI nr. 478*. Springer Verlag, Berlin, 1990.

[Bierwisch, 1982] M. Bierwisch. Semantische und konzeptuelle Repräsentation lexikalischer Einheiten. In R. Ružička and W. Motsch (eds.), *Untersuchungen zur Semantik*. Berlin, 1982.

[Boguraev and Pustejovsky, 1996] B. K. Boguraev and James Pustejovsky. *Corpus Processing for Lexical Acquistion*. MIT Press, Cambridge, MA, 1996.

[Bos *et al.*, 1995] J. Bos, P.P. Buitelaar and A.M. Mineur. Bridging as Coercive Accommodation. In S. Manandhar (ed.), *Proceedings of CLNLP95*, Edinburgh, 1995.

[Brent, 1993] Michael R. Brent. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19(2), 1993.

[Bresnan, 1982] Joan Bresnan. *The Mental Representation of Grammatical Representations*. MIT Press, Cambridge, MA, 1982.

[Brill, 1992] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*. ACL, 1992.

[Brown *et al.*, 1992] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai and Robert Mercer. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4), 1992.

[Brugman, 1981] Claudia Brugman. Story of *over*. Master's thesis, University of California, Berkeley, 1981.

[Buitelaar and Mineur, 1994] P. P. Buitelaar and A. M. Mineur. Coercion and Compositionality in Categorial Grammar. In Paul Dekker and Martin Stokhof (eds.), *Proceedings of the 9th Amsterdam Colloquium*, pages 175–188, Amsterdam, 1994. ILLC.

[Busa, 1997] Federica Busa. The Semantics of Agentive Nominals in the Generative Lexicon. In Patrick St.Dizier (ed.), *Predicative Forms in Natural Language*. Kluwer, the Netherlands, 1997.

[Calzolari, 1991] Nicoletta Calzolari. Representation of Semantic Information in a Lexical Knowledge Base. In E. Briscoe and A. Copestake (eds.), *First ACQUILEX Workshop*, Cambridge, MA, 1991.

[Cardelli and Wegner, 1985] L. Cardelli and P. Wegner. On Understanding Types, Data Abstraction, and Polymorphism. *Computing Surveys*, 17:22–29, 1985.

[Charniak, 1993] Eugene Charniak. *Statistical Language Learning.* MIT Press, 1993.

[Church and Hanks, 1990] K. W. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16:22–29, 1990.

[Copestake and Briscoe, 1993] A. Copestake and E. Briscoe. Lexical Operations in a Unification-Based Framework. In James Pustejovsky and S. Bergler (eds.), *Lexical Semantics and Knowledge Representation. Lecture Notes in Artificial Intelligence 627*, pages 22–29, Berlin, 1993. Springer Verlag.

[Copestake and Briscoe, 1995] A. Copestake and E. Briscoe. Semi-Productive Polysemy and Sense Extension. *Journal of Semantics*, 12(2), 1995.

[Cowie *et al.*, 1993] Jim Cowie, Louise Guthrie, James Pustejovsky, Takahiro Wakao, Jin Wang and Scott Waterman. The DIDEROT information extraction system. In *Proceedings of the First Conference of the Pacific Association for Computational Linguistics*, Vancouver, Canada, 1993.

[Cruse, 1986] D. A. Cruse. *Lexical Semantics.* Cambridge University Press, Cambridge, UK, 1986.

[Dagan *et al.*, 1997] Ido Dagan, Lillian Lee and Fernando Pereira. Similarity-Based Methods for Word Sense Disambiguation. In *Proceedings of ACL/EACL97*, Madrid, Spain, 1997.

[Dölling, 1994] Johannes Dölling. Sortale Selektionsbeschränkungen und Systematische Bedeutungsvariationen. In M. Schwarz (ed.), *Kognitive Semantik / Cognitive Semantics*, pages 167–216. Gunter Narr, Tübingen, 1994.

[Evans and Gazdar, 1996] Roger Evans and Gerald Gazdar. DATR: A Language for Lexical Knowledge Representation. *Computational Linguistics*, 22(2):167–216, 1996.

[Fillenbaum and Jones, 1965] S. Fillenbaum and L.V. Jones. Grammatical Contingencies in Word Association. *Journal of Verbal Learning and Verbal Behavior*, (4), 1965.

[Fillmore, 1968] Ch. Fillmore. The case for case. In E. Bach and R. Harms (eds.), *Universals in Linguistic Theory*. 1968.

[Gale *et al.*, 1992] W. K. Gale, K. Church and D. Yarowsky. One Sense per Discourse. 1992.

[Grefenstette, 1994] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston, 1994.

[Grefenstette, 1996] Gregory Grefenstette. Evaluating Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches. In B. Boguraev and J. Pustejovsky (eds.), *Corpus Processing for Lexical Acquistion*. MIT Press, 1996.

[Hindle, 1990] Donald Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the ACL*, pages 268–275, 1990.

[Hobbs *et al.*, 1993] J. Hobbs, M. Stickel, P. Martin and D. Edwards. Interpretation as Abduction. *Artificial Intelligence*, 63, 1993.

[Hobbs *et al.*, 1997] J. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel and M. Tyson. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In E. Roche and Y. Schabes (eds.), *Finite State devices for Natural Language Processing*, pages 383–406. MIT Press, Cambridge, MA, 1997.

[Johnston *et al.*, 1995] Michael Johnston, Bran Boguraev and James Pustejovsky. The Acquisition and Interpretation of Complex Nominals. In *Working Notes of AAAI Spring Symposium on the Representation and Acquisition of Lexical Knowledge*, Stanford University, 1995.

[Kamp and Reyle, 1993] Hans Kamp and Uwe Reyle. *From Discourse to Logic; introduction to modeltheoretical semantics of natural language, formal logic, and discourse representation theory.* Kluwer, 1993.

[Kayser, 1988] D. Kayser. What kind of a thing is a concept? *Computational Intelligence*, 4(2):158–165, 1988.

[Kilgarriff, 1992] Adam Kilgarriff. *Polysemy.* PhD thesis, University of Sussex, Brighton, UK, 1992.

[Kilgarriff, 1995] Adam Kilgarriff. Inheriting Polysemy. In Patrick Saint-Dizier and Evelyne Viegas (eds.), *Computational Lexical Semantics.* Cambridge University Press, 1995.

[Kilgarriff, 1997a] Adam Kilgarriff. Evaluation of word sense disambiguation programs: progress report. In *Proceedings of SALT Workshop on Evaluation in Speech and Language Technology.* Sheffield University, UK, 1997.

[Kilgarriff, 1997b] Adam Kilgarriff. Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora. 1997.

[Lakoff, 1987] George Lakoff. *Women, Fire and Dangerous Things.* University of Chicago Press, Chicago, IL, 1987.

[Lascarides and Copestake, 1995] Alex Lascarides and Ann Copestake. The Pragmatics of Word Meaning. In *Semantics and Linguistic Theory*, 1995.

[Leiss, 1990] Hans Leiss. Polymorphic Constructs in Natural and Programming Languages. In J. van Eyck (ed.), *JELIA90 Logic in AI. Lecture Notes in Artificial Intelligence 478.* Springer Verlag, Berlin, 1990.

[Levin, 1993] Beth Levin. *English Verb Classes and Alternations.* University of Chicago Press, 1993.

[Luelsdorff, 1994] Philip A. Luelsdorff. *The Prague School of structural and functional linguistics.* Linguistic and Literary Studies in Eastern Europe 41. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1994.

[Martin, 1991] James H. Martin. Representing and Acquiring Metaphor-Based Polysemy. In Uri Zernik (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Buid a Lexicon.* Lawrence Erlbaum Ass., Hillsdale, NJ, 1991.

[Mercer, 1993] Robert E. Mercer. Presuppostions and Default Reasoning: A Study in Lexical Pragmatics. In James Pustejovsky and S. Bergler (eds.), *Lexical Semantics and Knowledge Representation. Lecture Notes in Artificial Intelligence 627.* Springer Verlag, Berlin, 1993.

[Miller *et al.*, 1990] G. A. Miller, R. Beckwith, Ch. Fellbaum, D. Gross and K. Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3,4, 1990.

[Mineur and Buitelaar, 1996] A. M. Mineur and P. P. Buitelaar. A Compositional Treatment of Polysemous Arguments in Categorial Grammar. In Kees van Deemter and Stanley Peters (eds.), *Semantic Ambiguity and Underspecification*, pages 125–143. CSLI-Chicago University Press, Chicago, 1996.

[Moens and Steedman, 1987] M. Moens and M. Steedman. Temporal Ontology in Natural Language. In *Proceedings ACL*, 1987.

[Montague, 1974] Richard Montague. Towards a Proper Treatment of Quantification in English. In R. H. Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague.* Yale University Press, New Haven, CT, 1974.

[Moortgat, 1988] M. Moortgat. *Categorial Investigations. Logical and Linguistic Aspects of the Lambek Calculus*, Band 9 der *Groningen-Amsterdam Studies in Semantics.* Foris, Dordrecht, 1988.

[Morrill, 1992] Glyn Morrill. Type-logical Grammar. Technical Report OTS-WP-CL-92-002, Onderzoeksinstituut voor Taal en Spraak, Utrecht, 1992.

[Nirenburg and Raskin, 1987] S. Nirenburg and V. Raskin. The Subworld Concept Lexicon and the Lexicon Management System. *Computational Linguistics*, 13:276–289, 1987.

[Nunberg, 1979] Geoffrey Nunberg. The Non-uniqueness of Semantic Solutions: Polysemy. *Linguistics and Philosophy*, 3, 1979.

[Ostler and Atkins, 1993] N. Ostler and B. Atkins. Predictable meaning shift: some linguistic properties of lexical implication rules. In James Pustejovsky and S. Bergler (eds.), *Lexical Semantics and Knowledge Representation. Lecture Notes in Artificial Intelligence 627*, Berlin, 1993. Springer Verlag.

[Partee *et al.*, 1993] Barbara Partee, Alice ter Meulen and Robert E. Wall. *Mathematical Methods in Linguistics.* Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.

[Partee, 1987] Barbara Partee. Noun Phrase Interpretation and Polymorphic Types. In Jeroen Groenendijk, Dick de Jongh and Martin Stokhof (eds.), *Studies in Discourse Representation Theory and the Theory of Generalized Quantifiers.* Foris, Dordrecht, the Netherlands, 1987.

[Pereira *et al.*, 1993] Fernando Pereira, Naftali Tishby and Lillian Lee. Distributional Clustering of English Words. In *Proceedings of ACL93*, pages 183–190, 1993.

[Poesio and Vieira, 1998] Massimo Poesio and Renata Vieira. A Corpus-Based Investigation of Definite description Use. *Computational Linguistics*, 24, 1998.

[Pollard and Sag, 1994] Carl Pollard and Ivan A. Sag. *HPSG.* CSLI-University of Chicago Press, 1994.

[Porter, 1980] M.F. Porter. An Algorithm for Suffix Stripping. *Program*, 3(14):130–137, 1980.

[Pustejovsky and Boguraev, 1993] James Pustejovsky and Bran Boguraev. Lexical Knowledge Representation and Natural Language Processing. *Artificial Intelligence*, 63:193–223, 1993.

[Pustejovsky and Busa, 1995] James Pustejovsky and Federica Busa. Unaccusativity and Event Composition. In P-M. Bertinetto, V. Binachi, J. Higginbotham and M. Squartini (eds.), *Temporal Reference: Aspect and Actionality.* Rosenberg and Sellier, Turin, Italy, 1995.

[Pustejovsky *et al.*, 1993] James Pustejovsky, Sabine Bergler and Peter Anick. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics*, 19(2), 1993.

[Pustejovsky *et al.*, 1995] James Pustejovsky, Bran Boguraev and Michael Johnston. A Core Lexical Engine: The Contextual Determination of Word Sense. Technical Report, Department of Computer Science, Brandeis University, 1995.

[Pustejovsky, 1989] James Pustejovsky. Current Issues in Computational Lexical Semantics. In *Proceedings ACL89*, pages xvii–xxv, 1989.

[Pustejovsky, 1991a] James Pustejovsky. The Generative Lexicon. *Computational Linguistics*, 17(4), 1991.

[Pustejovsky, 1991b] James Pustejovsky. The syntax of event structure. *Cognition*, 41:47–81, 1991.

[Pustejovsky, 1994a] James Pustejovsky. Linguistic Constraints on Type Coercion. In P. Saint-Dizier and E. Viegas (eds.), *Computational Lexical Semantics*. Cambridge University Press, 1994.

[Pustejovsky, 1994b] James Pustejovsky. Semantic Typing and Degrees of Polymorphism. In Martin-Vide (ed.), *Current Issues in Mathematical linguistics*. Elsevier, 1994.

[Pustejovsky, 1995] James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

[Pustejovsky, 1996] James Pustejovsky. The Semantics of Complex Types. *Lingua*, 1996.

[Resnik, 1992] P. Resnik. Wordnet and Distributional Analysis: A Class-Based Approach to Lexical Discovery. In *Workshop Notes, Statistically-Based NLP Techniques*, pages 54–64. AAAI, 1992.

[Ruhl, 1989] C. Ruhl. *On Monosemy: A study in linguistic semantics*. State University of New York Press, Albany, NY, 1989.

[Sanfilippo, 1993] Antonio Sanfilippo. LKB Encoding of Lexical Knowledge from Machine-Readable Dictionaries. In A. Copestake, E. Briscoe and V. de Paiva (eds.), *Inheritance, Defaults and the Lexicon*, pages 190–222. Cambridge University Press, Cambridge, UK, 1993.

[Schütze, 1992] Hinrich Schütze. Word Sense Disambiguation with Sublexical Representations. In C. Weir, S. Abney, R. Grishman and R. Weischedel (eds.), *Workshop Notes, Statistically-Based NLP Techniques*, pages 109–113. AAAI, 1992.

[Schütze, 1994] Hinrich Schütze. Towards Connectionist Lexical Semantics. In Susan D. Lima, Roberta L. Corrigan and Gregory K. Iverson (eds.), *The reality of Linguistic Rules*, Studies in Language Companion Series 26, pages 171–191. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1994.

[Shieber, 1992] S. Shieber. *Constraint-Based Grammar Formalisms. Parsing and Type Inference for Natural and Computer Languages.* MIT Press, Cambridge, MA, 1992.

[Smadja, 1993] Frank Smadja. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1), 1993.

[Sowa, 1984] J. F. Sowa. *Conceptual structures: Information processing in mind and machine.* Addison Wesley, 1984.

[Sowa, 1993] J. F. Sowa. Logical Structures in the Lexicon. In James Pustejovsky and S. Bergler (eds.), *Lexical Semantics and Knowledge Representation. Lecture Notes in Artificial Intelligence 627*, pages 577–613, Berlin, 1993. Springer Verlag.

[Stern, 1931] Gustaf Stern. *Meaning and Change of Meaning.* Indiana University Press, Bloomington, IN, 1931. Reprint, 1964.

[Tesniere, 1959] Lucien Tesniere. *Elements de Syntaxe Structurale.* Librairie C. Klincksieck, Paris, 1959.

[Touretzky, 1986] D. Touretzky. *The Mathematics of Inheritance Systems.* Morgan Kaufmann, Los Altos, CA, 1986.

[van Deemter and Peters, 1996] Kees van Deemter and Stanley Peters. *Semantic Ambiguity and Underspecification.* CSLI, Stanford, CA, 1996.

[Verhagen, 1996] Marc Verhagen. ms. Brandeis University. Waltham, MA, 1996.

[Verschuur, 1993] Leon Verschuur. Flexibly Disambiguating Metonymy. Technical Report MMC 52, ITK, Tilburg, 1993.

[Verspoor, 1996] Karin Verspoor. Lexical Limits on the Influence of Context. In G. W. Cottrell (ed.), *Proceedings Cogsci96*. Lawrence Erlbaum Associates, 1996.

[Viegas *et al.*, 1996] Evelyne Viegas, Boyan Onyshkevych, Victor Raskin and Sergei Nirenburg. From *Submit* to *Submitted* via *Submission*: On Lexical Rules in Large-Scale Lexicon Acquisition. In *Proceedings ACL96*, pages 32–39. ACL, 1996.

[Weinreich, 1964] Uriel Weinreich. Webster's Third: A Critique of its Semantics. *International Journal of American Linguistics*, 30:405–409, 1964.

[Weinreich, 196x] Uriel Weinreich. Dictionaries of the Future: A Set of Parameters for Descriptive Semantics. 196x. First published in: Uriel Weinreich on Semantics (W. Labov and B.S. Weinreich editors).

[Wilks and Stevenson, 1996] Yorick Wilks and Mark Stevenson. The Grammar of Sense: Is word-sense tagging much more than part-of-speech tagging? Technical Report CS-96-05, University of Sheffield, UK, 1996.

[Yarowsky, 1992] D. Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, 1992.

[Zernik, 1991] U. Zernik. *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

Appendix A

CoreLex **Type Index**

This is a complete list of all 126 CoreLex underspecified semantic types. For an overview that includes all instances for each type see the CoreLex webpage at:

`http://www.cs.brandeis.edu/~paulb/CoreLex/overview.html`

The types are ordered alphabetically on their three letter codes. Each code is followed by the number of instances for this type and by a list of basic types, to which the underspecified semantic type corresponds. The total number of nouns that are included in the CoreLex database is 39,937.

abs	2	abstraction
acp	115	act attribute process psychological_feature state
acr	27	act attribute event relation state
acs	93	act state
act	2,999	act
aes	19	act event state
aev	133	act event
age	6	act causal_agent
agh	28	causal_agent human
agl	6	causal_agent location
agm	8	causal_agent animal
agp	2	causal_agent psychological_feature
agt	468	causal_agent
anf	104	animal food
anm	1,947	animal
ann	19	animal artifact natural_object
anp	7	animal psychological_feature
aqu	42	artifact quantity_definite quantity_indefinite
ara	39	artifact attribute psychological_feature state
arg	17	artifact group
arh	165	artifact human
arp	54	artifact psychological_feature state
art	4,644	artifact state
atc	61	attribute communication phenomenon psychological_feature state
ate	202	act attribute event psychological_feature state
atg	5	attribute group
atl	12	attribute location psychological_feature state
atp	273	attribute psychological_feature state
atr	2,376	attribute process state
avf	4	act artifact form
avl	10	act artifact location location_geographical
avr	10	act artifact event relation
avt	292	act artifact attribute event physical_object process state
caa	4	artifact attribute communication
cae	179	act artifact communication psychological_feature state
cea	88	act attribute communication event psychological_feature state
cel	48	cell
cha	20	artifact chemical substance
chf	5	chemical food
chm	570	chemical state
chp	11	chemical plant
coa	303	act communication event process psychological_feature relation state
coh	110	communication human
col	7	communication location_geographical
com	1,760	communication psychological_feature relation state
con	10	act consequence event
ent	7	entity
evs	31	event process state
evt	345	event
fac	15	act artifact food

fev	2	food group_social
fod	825	attribute food form natural_object quantity_indefinite state substance
frc	9	communication form psychological_feature
fre	13	act event form state
frm	144	attribute form state
frt	43	artifact attribute form location natural_object state
ftp	9	attribute food plant substance
gas	52	act communication event group_social psychological_feature state
gbp	6	group_biological psychological_feature
gbs	7	group_biological group_social
grb	4,798	group_biological
grp	127	group
grq	8	group_social quantity_definite
grs	323	group group_social state
gsa	83	act artifact attribute communication group group_social substance
gsl	20	group_social location_geographical
gsp	45	group_social psychological_feature state
hue	62	act event human state
hum	4,523	group group_social human natural_object
hup	34	human psychological_feature
lac	16	act location location_geographical
lap	56	artifact communication location psychological_feature state
lfr	46	life_form
lgg	5	group location_geographical
lme	61	linear_measure quantity_definite
loc	255	artifact location location_geographical possession state
log	997	artifact location_geographical natural_object
lon	12	location natural_object
lor	7	location_geographical relation
mea	21	measure
mic	162	microorganism
naa	52	artifact natural_object
nac	8	act artifact event natural_object
naf	7	form natural_object
naq	2	natural_object quantity_indefinite
nat	683	natural_object substance
pan	104	artifact natural_object plant substance
pap	3	act attribute possession psychological_feature
pas	20	act attribute possession relation state
pgb	10	group_biological plant
phm	199	act event phenomenon process relation state
pho	102	physical_object
php	9	event phenomenon psychological_feature
phs	2	phenomenon substance
pht	12	attribute phenomenon
plf	214	food plant
plt	1,592	food natural_object plant substance
poa	72	act communication event possession psychological_feature state
pom	26	act artifact attribute communication possession
poq	7	possession quantity_definite

pos	227	**possession process psychological_feature state**
prn	3	**natural_object process**
pro	422	**process**
prt	651	**artifact attribute food form human natural_object part quantity_indefinite state substance**
psa	28	**act artifact possession**
psg	17	**act group psychological_feature**
psr	2	**possession relation**
psy	1,604	**event psychological_feature state**
pya	209	**act event psychological_feature relation state**
qcc	8	**quantity_definite state**
qcs	4	**communication quantity_definite state**
qde	5	**act quantity_definite**
qie	9	**act event quantity_indefinite**
qud	466	**quantity_definite**
qui	134	**quantity_indefinite**
reg	6	**group relation**
rel	317	**attribute human psychological_feature relation state**
saa	71	**artifact attribute substance**
sas	15	**act state substance**
spc	8	**space**
spe	2	**act space**
sta	1,003	**state**
sub	733	**substance**
tme	355	**group phenomenon quantity_definite state time**
tmt	5	**attribute time**
tmv	36	**act event time**

Appendix B

Program Descriptions

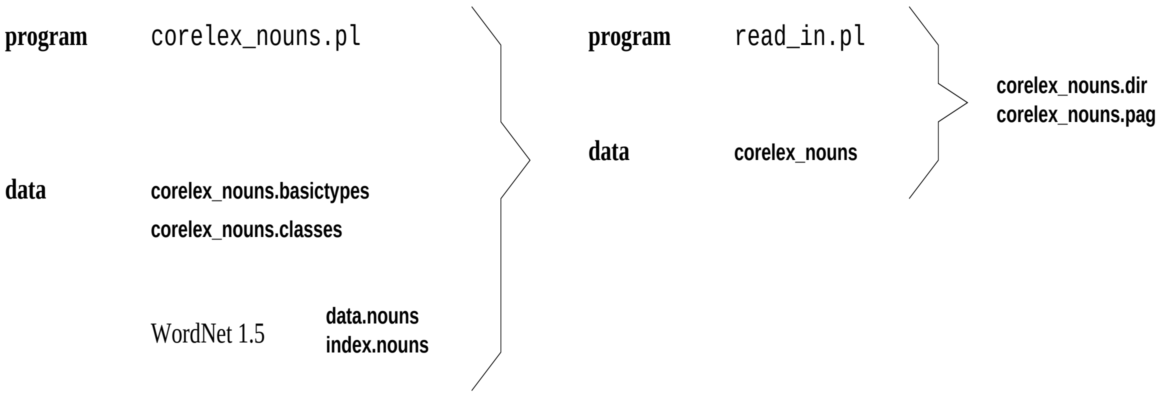

Figure B.1: Input and output for `corelex_nouns.pl`

The program `corelex_nouns.pl` is used to generate the CoreLex database (**corelex_nouns**)[1] from the WordNet databases (**data.nouns** and **index.nouns**). The basic types in CoreLex as defined in Section 4.2 are stored in **corelex_nouns.basictypes**. The combinations of such types, as they appear in WordNet, and how they correspond to the underspecified semantic types in CoreLex, are stored in **corelex_nouns.classes**. These files are the input for the program. The output is the CoreLex database, which is subsequently stored as a PERL database file (**corelex_nouns.dir** and **corelex_nouns.pag**) by the program `read_in.pl`.

[1] That is, for nouns, although similar programs were constructed also for adjectives and verbs, but contrary to the data on nouns, the resulting data of those programs has not been analysed manually, as was discussed in Section 5.4.

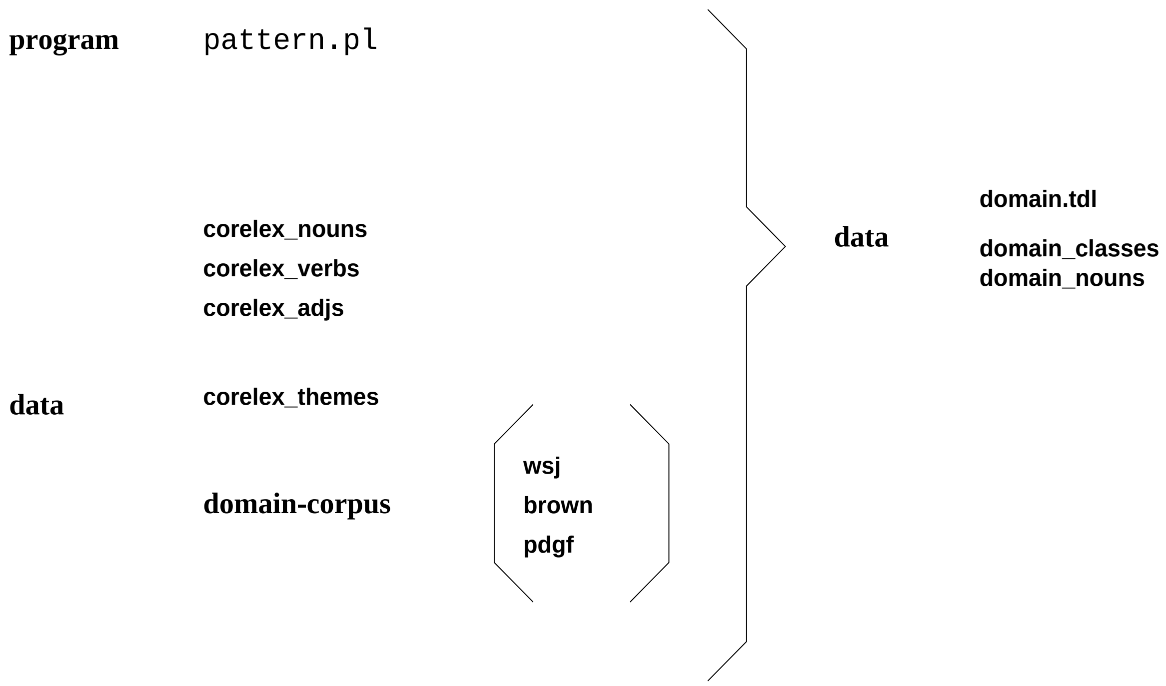

Figure B.2: Input and output for `pattern.pl`

The program `pattern.pl` is used for generating a 'domain database', that is, a database of nouns (**domain_nouns**) in a large corpus that is representative of a particular domain, supplied with information on the frequency and mutual information scores of the linguistic patterns in which these nouns occur in the corpus, as discussed in Chapter 5. The same information is also grouped according to the types of the nouns (somewhat incorrectly named **domain_classes**). Additionally a file is generated that restructures, again, the same information in a qualia structure style format (**domain.tdl**).

These output files are generated on the basis of several inputs: a particular corpus (**wsj, brown, pdgf**), the CoreLex database for nouns (**corelex_nouns**) – plus additionally for verbs and adjectives (**corelex_verbs** and **corelex_adjectives**) –, and a database of prepositions, CoreLex types and thematic roles (**corelex_themes**), that is used for the assignment of functional structure as discussed in Section 5.3.

The domain databases generated by `pattern.pl` are used by the program `class.pl` to generate a classified version (**text.classified**) of a given input text (e.g. **stocks, nren, orwell, mr, darwin**).

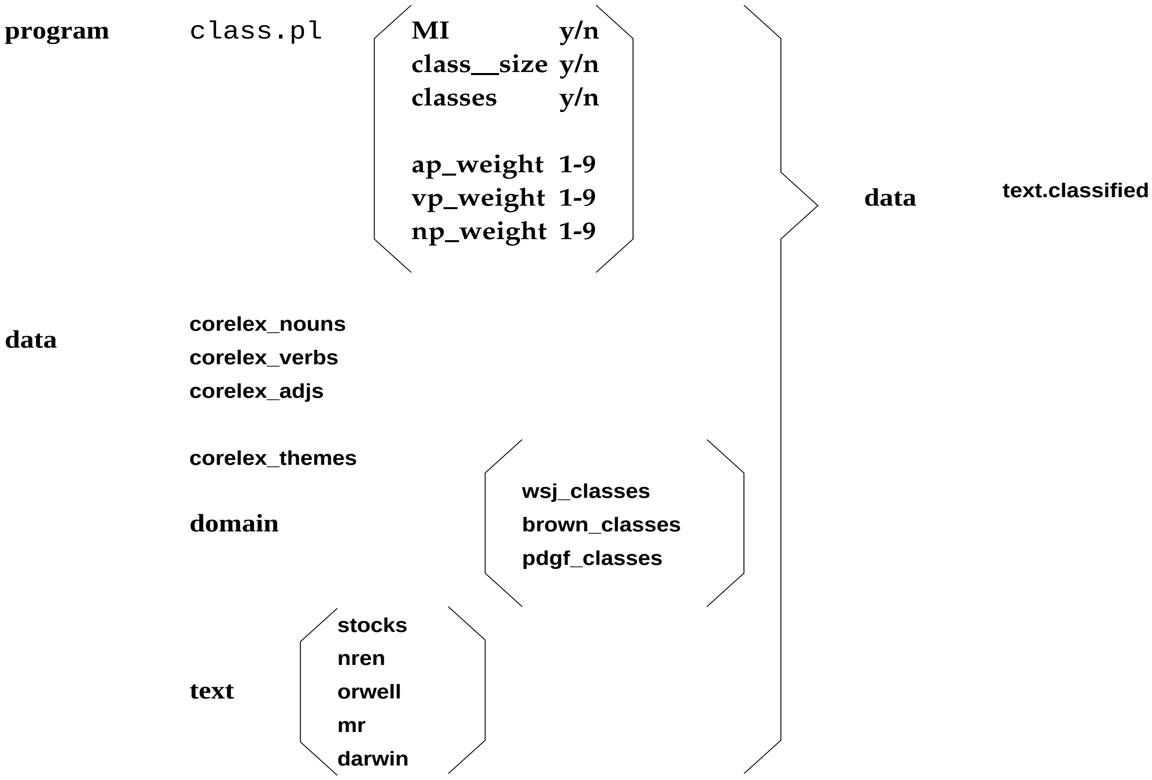

Figure B.3: Input and output for `class.pl`

This includes classifying unknown nouns (those that are not in CoreLex) and tagging nouns with a context dependent interpretation. Apart from the text and domain database inputs, other inputs are identical to those for `pattern.pl`. The parameters on classification that were discussed in Section 6.4.3 are three boolean choices and three weights. The booleans are on classification using mutual information or frequency numbers (MI), on classification using relative class size (class_size) and on classification using lexical items or classes of lexical items (classes). The weights are on different phrase structures (ap_weight, vp_weight, np_weight).