Order Number 9510717

The treatment of phrasal verbs in a natural language processing system

Shaked, Nava Ayala, Ph.D.

City University of New York, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

THE TREATMENT OF PHRASAL VERBS IN A NATURAL LANGUAGE PROCESSING SYSTEM

by

Nava Ayala Shaked

A dissertation submitted to the Graduate Faculty in Linguistics
In partial fulfillment of the requirements for the degree
Doctor of Philosophy, The City University of New York.

1994

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

4 September 1994 — Date
Virginia Teller — Chair of the Examining Committee

September 6, 1994 — Date
Charles E. Cairns — Executive Officer

Professor Martin Chodorow

Professor William Stewart

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE TREATMENT OF PHRASAL VERBS IN A NATURAL LANGUAGE PROCESSING SYSTEM

by

Nava Ayala Shaked

Advisor: Professor Virginia Teller

The phrasal verbs construction is problematic both theoretical and computational linguistics. The problems involve all areas from representation to processing. This work addresses the treatment of phrasal verbs in a computational lexicon. The purpose is to arrive at an adequate classification of the construction that will minimize its ambiguity and conform to the needs of both theoretical and computational linguistics.

While in theoretical linguistics there is extensive literature on the issue, there are few attempts to deal with the construction in the area of computational linguistics. Current treatments of phrasal verbs suggested by NLP systems tend to be ad hoc and lack both systematic classification and theoretical support. This research offers a detailed look at this ambiguous construction, highlights some of the problems and offers solutions. We claim that the key to successful processing lies in the correct representation of phrasal verbs which require different internal classification. This can be done based on the syntactic and semantic attributes of the construction coupled with some non-linguistic information such as statistical information.

The model presented in this work accounts for the lexical, structural, and semantic representation of the three types of phrasal verbs: literal, systematic, and figurative. The

model assumes a dependency between the verb and the particle, and uses semantic features in order to capture the different possible senses a phrasal verb can have. We also present some initial implementation and testing of the model for the task of assigning lexical ambiguity tags.

The treatment of phrasal verb is relevant to the treatment of other constructions which pose a problem for both linguistic theory and computational systems. These include compounds, idioms in general, and long distance relationships similar to those exhibited by verbs and particles. The three-level model sheds light on possible treatments in these problematic areas as well.

Acknowledgements

To put together this dissertation took a lot of time, effort, support, and love. It was the most challenging thing I have ever experienced, and it was possible to do thanks to the people who believed in me and despite those who did not.

I would not be here in New York if it weren't for the encouragement and support of my undergraduate mentors, Malka Rappaport Hovav and Susan Rothstein, who taught me the basics of linguistics and encouraged me to continue and explore.

It was here at CUNY that John Moyne introduced me to the area of computational linguistics and helped me pursue it in my graduate work, for that I am immensely grateful.

To my committee members, Martin Chodorow and Bill Stewart, who read, evaluated, and commented on the work; and most of all, to Virginia Teller, who stuck by me and believed in me all along, spending hours on weekends accommodating my emotional and intellectual needs my heartfelt thanks. Virginia, you have been a mentor, a friend, and an exemplary adviser.

Many thanks, also, to my friends in the CUNY Linguistics Program: Beth Craig ,Rick McKinnon, Robert Hollander, Ivy Sichell, Iris Elisha, Barbara Bevington, Harriet Taber, Mirianna Washburn, and many more who were on the same path for the last six years. A special thanks to my friend Nicholas Papacostas, whose friendship and kindness were always appreciated.

Also thanks to Judith Tucker for being there for all of us in the department and for assisting me in editing and proofreading this work.

Thanks to all my colleagues at AT&T Bell Labs who opened the magic world of computational linguistics to me and taught me wonderful things about speech and language. Joe Olive and Julia Hirschberg gave me the opportunity to work for them for two years, Ken Church, Don Hindle, and David Yarowski, helped me with my ACL paper. To Ido Dagan for long talks, a place to crash some nights, and much more, and above all to my dearest friend Jill Burstein, who was always there for me, keeping my spirits up, especially during the hard times of pregnancy.

Also thanks to Ken Ross from Boston University for providing me with the speech files and analysis on which section 2.2.1.2 is based.

To my colleagues and friends at NYNEX Science & Technology, which has my home for the last two years, thanks for giving me challenging work, good company, and lots of support. Many thanks to Mike Cohen for letting me use his facilities and for helping me organize this work. Most of all, thank you Sara Basson for always being there for me professionally and personally, and thanks also for many long discussions about priorities, choices, and parenthood, and for helping me sort things out logically.

To Phyllis, thanks for many hours of talks and encouragement and for a very unique friendship.

To our friends who spent days entertaining Udi and Britt while I was studying, who opened thier hearts and their houses, thank you Amir and Penina, Rami and Tami, Zafrir and Lisa and their children.

Thanks to Ricky Tadmor for giving me the opportunity to be part of Israeli education in "Oranim" and for being a real friend in all senses of the word and to all my students in

Oranim who helped me keep in touch with Israel and made me feel that I have made a difference.

Both our families, although away, tried to make a difference. Doris and Abraham, thanks for sending all those delicious goodies from home and thinking of us always. Thanks to my parents who came to help in critical times and for a childhood lesson that everything is possible if you really want it. Thanks to my cousins Zehava and Gadi Alon for endless conversations and advice for both of us. But above all thanks to the Rons: Yael, Gadi, Shany, Topaz and Orel. Your endless letters, attention, and support kept us alive and made us feel loved and close to the family through difficult times. We will never be able to repay such a debt.

To my daughter, Britt Eti who was the guiding light and the source of endless happiness and love during the last three years. Thanks for being so cooperative and understanding during my long hours of studying and the days I was gone. I love you.

To Shosh and the whole Ben-Yaakov family, who provided Britt with a second home while taking care of her days and nights and were always there for Udi and me. And finally to my husband Udi, who kept pulling the wagon when it almost looked hopeless. Thanks for sharing with me and being both a mother and a father to Britt. Without your help this would have never been possible.

Finally, thanks to all those who crossed my path during these past six years and helped shape my work and thought.

Table of Contents

List of Tables

CHAPTER I: INTRODUCTION

1.1 MAJOR QUESTIONS AND HYPOTHESES

This work addresses the treatment of the phrasal verb construction in natural language processing (NLP) systems. Since ambiguity is the most pervasive phenomenon in natural language processing and must be handled by any NLP system, ambiguity resolution is crucial for language analysis, understanding, and interpretation. An NLP model typically includes three main components: a lexical component for part of speech assignment, a structural component for syntactic assignment and a semantic component for interpretation. Because of the multiple ambiguities that the phrasal verb construction displays, its correct representation in the computational lexicon is crucial for lexical, structural, and semantic analysis. In addition, with the appropriate lexical representation, a great deal of the ambiguity can be resolved at the lexical level rather then carrying it to subsequent processing levels.

The first issue that this work addresses is whether phrasal verbs can be defined and classified in a way that will reduce their complexity and resolve their ambiguity, while satisfying the requirements of both theoretical and computational linguistic research.

A second issue involves the internal components of the phrasal verb. Do we need a special part of speech PARTICLE (PART), or can we collapse these occurrences under

Preposition (PP) or under Adverbials (Adv) as proposed by several scholars (Aarts 1989, Radford 1988, among others). In what follows I present phonological, syntactic, morphological, and semantic justifications of the claim that the phrasal verb is a combination of a verb and a particle. Resolving this second issue is crucial to the definition and classification of phrasal verbs. This work aims to show that in order to define and classify phrasal verbs we cannot rely on either a structural or a semantic description alone, but must construct a combination of the two. To arrive at an appropriate definition and classification, we must consider the relationship between the verb and the particle, rather than examining the particle or the verb independently.

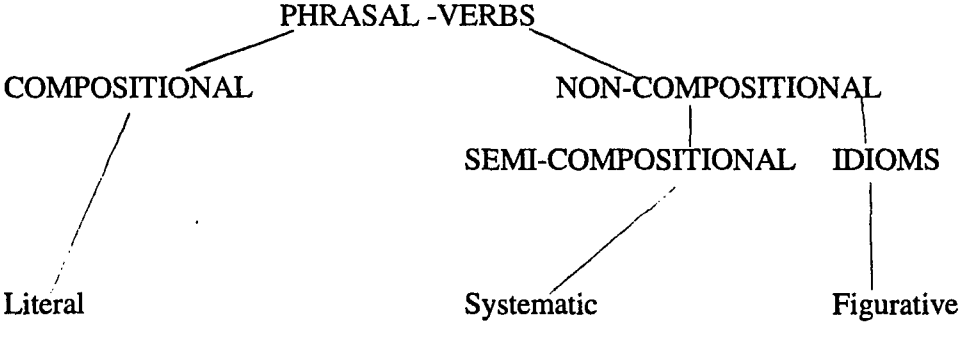
The final issue addressed involves the relationship between stochastic and linguistic approaches to NLP. This thesis argues that a combination of the two approaches will lead to better performance than either approach alone. An experiment has been conducted to evaluate the performance of PARTS, a stochastic lexical analyzer for the task of tagging phrasal verbs with their correct part of speech assignments. The results indicate that enhancing the stochastic model with linguistic information reduced the error rate by 7% for this task.

The phrasal verb construction is interesting from the perspective of both computational and theoretical linguistics. The lexical, structural, and semantic status of phrasal verbs remain active topics of debate, with disagreements focusing on the dual behavior of the phrasal verb as one or two constituents, the different semantic levels, and the long distance relationship between the discontinuous verb and particle.

For computational linguistics (CL), resolution of these issues is even more crucial

than for theoretical linguistics. In a computational system, one cannot depend on human intuition to make subtle semantic and structural judgments; instead one must provide a detailed representation that will enable successful processing.

Another important representation debate in the literature is whether the verb and particle should appear as one or two units (in the computational lexicon). Because of the non-compositional semantic interpretation of some phrasal verbs, some linguists regard them as idioms that should be put in a list separate from the lexicon. This thesis distinguishes between phrasal verbs of compositional meaning and combinations of non-compositional meaning. According to Fraser (1976), "literal" phrasal verbs are considered to be compositional combinations while "systematic" and "figurative" phrasal verbs are classified as non-compositional. Non-compositional combinations are further divided into semi-compositional and idioms, as shown below:



I will argue that "literal" phrasal verbs do not require any special treatment in the computational lexicon. Due to the non-compositional nature of "systematic" phrasal verbs, additional information about the relationship between the verb and the particle is required

for interpretation purposes. Nonetheless, this can still be handled in the regular lexicon. "Figurative" combinations, however, should be represented in a special list. This work describes a set of computational procedures (the CEC tool) that determine how such a list should be formed for the stochastic tagger PARTS.

The main issues and hypotheses of this thesis can be summarized as follows. The main purpose is to arrive at an adequate classification of phrasal verbs that will help to minimize their ambiguity and will conform to the needs of both theoretical and computational linguistics. This ambiguity is reinforced by the dual behavior of the phrasal verb, sometimes as one complex verb and sometimes as two separate lexical items.

In order to do this the first hypothesis is that the phrasal verb is a combination of a verb and a particle and that an appropriate representation and classification must consider the syntactic and the semantic relationship between the verb and the particle in tandem rather then considering each separately.

Finally, this work advocates the use of both stochastic and linguistic information for the purposes of identifying, disambiguating, and tagging phrasal verbs.

The next section describes the five chapters of this thesis.

1.2 DESCRIPTION OF THE CHAPTERS

The phrasal verb construction is a problematic construction to both theoretical and computational linguistics. The problems involve all areas from representation to processing. It will be argued that in order to analyze or represent phrasal verbs in the computational lexicon one should consider the relationship between the verb and the particle, rather than simply deciding whether they should be considered as one unit or as two separate units. In addition, the processing of phrasal verbs in the stochastic tagger PARTS is described and several ways to improve the tagger's performance are suggested. The solution involves enhancing the statistical tagger with linguistic knowledge.

Chapter I outlines the major questions and hypotheses addressed. One main problem is the definition of phrasal verbs. Section 1.3.2 discusses the various definitions and tests proposed in the linguistic literature, pointing out that no one test actually applies to all types of phrasal verbs. Each study defines the construction according to a limited range of criteria such that some are based on semantic criteria, some on syntactic criteria, and others on a mixture of both. In some studies certain types of phrasal verbs which pose a special problem are excluded. Section 1.3.3 defines phrasal verbs for the purposes of this research and presents the basic reasons for this choice.

Chapter II focuses on the problem of phrasal verb ambiguity. The introduction describes in general what types of ambiguity exist, and section 2.2 deals with the importance of resolving phrasal verb ambiguity from the point of view of an NLP system. Three main areas of potential ambiguity are discussed: part of speech tagging, syntactic structure, and semantic interpretation. For each area, the ambiguity problems are detailed and existing NLP

systems and their treatment of such ambiguity are described. This section also illustrates the consequences of an error in identification or processing the phrasal verb construction. The chapter ends by raising the various aspects that must be dealt with in describing how to correctly represent the construction in a computational lexicon, which cannot rely on human intuitions in making choices.

Chapter III focuses on the theoretical aspects of the phrasal verb construction, to show that many issues are still not resolved from the point of view of either theoretical linguistics or computational linguistics. The main problems are whether the verb and the particle should be regarded as one lexical unit or two, and whether the verb and the particle combination should be stored and generated in the morphological or the syntactic component of the grammar. Other issues that are still being debated include the semantic classification of phrasal verbs, as well as their phonological properties, and aspects of acquisition and processing. In this chapter different approaches are presented and the benefits of adopting one view over another for the purposes of NLP are considered.

Chapter IV proposes a three-level model of representation of phrasal verbs in a computational lexicon and presents the results of a partial test of this model using the stochastic tagger PARTS. The model resolves several problematic issues of representation and offers alternatives to existing treatments. The second part of chapter IV turns to implementation. An experiment previously described (Shaked 1993) shows how problematic the phrasal verb construction is for the stochastic tagger PARTS, and compares two types of changes designed to improve its performance. The enhancements incorporated in the three-level model proved superior. These modifications include considering some pairs as one

lexical unit and adding linguistic knowledge to the statistical algorithm. Together these changes reduced the error rate for identifying and correctly tagging phrasal verbs by 7%. The chapter ends with several examples of processing that show how the fully implemented model would correctly handle difficult cases of phrasal verb ambiguity.

Chapter V summarizes the conclusions of this research and discusses future work. It is clear that when dealing with phrasal verbs more emphasis should be put on relationship between the verb and the particle both the structural and the semantic. Furthermore, in order to be able to identify and to process phrasal verbs in an NLP lexicon a better representation is required, one that will accommodate the many-to-many relationship between the verb and the particle, yet at the same time will prevent similar verb- preposition structures from being included. The issue of verb+particle relationship is also related to long distance dependencies, compound structures, and the handling of idioms in general. Future tasks include applying the ideas proposed for phrasal verbs to other similar constructions, such as compounds and idioms in general.

1.3 THE DEFINITION OF PHRASAL VERBS

1.3.1. Introduction

Modern English grammar speaks of a special category of verbs or verbal constructions.

1. He looked up his friend in the phone book.

2. You're putting him on.

3. The plane took off on time.

Various names are used in the literature for this form: 'two-word verb', 'discontinuous verb', 'compound verb', 'verb particle constituent', 'verb adverb compound', 'verb particle combination' among others. All definitions refer to the phenomenon in which a verb and another element, which we will call a particle, are said to constitute a kind of integral unit. The relations between the verb and the particle in such constructions and the linguistic description of the combination vary among definitions and will be a major focus of investigation in this research.

McPartland (1987) claims that phrasal verbs are a subset of idioms, composed of a verb and an adverbial particle. For her, phrasal verbs are complex units with a non-compositional interpretation, yet are known to native speakers of the language. Bolinger (1971:12) describes the phrasal verb as "a lexical unit in the strict sense of non-additive compound or derivative, one that has a set meaning which is not the sum of the meanings of its parts." Kennedy (1920) and Makkai (1972) consider some verb adverb combinations as phrasal verbs, arguing that, despite the fact that their meaning is predictable, certain verbs

and adverbs show a strong tendency to co-occur and therefore constitute a semantic unit[1].

I will use the term "phrasal verbs", which was introduced by Smith (1925), to refer to this construction. Smith regarded the phrasal verb as an idiom expressing a single notion, where the meaning is not implied by the components and may correspond to one word in other languages. For example, *fall out* has parallel, single word counterparts in Latin *excideve* and in German *ausfallen*. For Smith the adverbial/preposition distinction is not relevant. The defining characteristic of phrasal verbs according to Smith is whether the verb together with the particle constitutes a semantic unit with a non compositional meaning. As Sroka (1972) points out, Smith emphasizes unity and irregularity.

Bolinger (1971) states that phrasal verbs can be defined by simply listing them. However, such a list will never be exhaustive, because new phrasal verbs are continually being added to the language. Bolinger also suggests that it might be feasible to list the particles. Although the particle class is no doubt smaller than the verb class, it is difficult to decide which words to include, because the definitions of phrasal verbs in the literature are often not clear or consistent on this point.

In the following section I will review the literature on the definition of phrasal verbs and will survey the different tests that have been suggested to determine which elements could qualify as phrasal verbs, in particular the question of what distinguishes the verb + particle combination from verb+preposition or verb+adverb. Section 1.1.2 defines what constitutes a phrasal verb for the purposes of this study.

---

[1]The co-occurrence of certain verbs with certain particle provides distributional evidence which some use to justify the one unit hypothesis.

1.3.2 General Definition: Literature Review

Pelli (1972) claims that defining phrasal verbs requires finding a linguistically tenable procedure that is as comprehensive and simple as possible in order to extract from the language a coherent and comparable group of verb particle constructions. Palmer (1968) states that the problem is to establish under what conditions such combinations are in some sense single units and suggests two types of classification - grammatical and semantic - that he insists must be kept distinct. This review describes the syntactic and the semantic criteria proposed by various researchers and shows that a definition of phrasal verbs cannot be achieved by using one or the other classification but rather features of both types.

There are two aspects to the definition of phrasal verb: one aspect deals with the distinction between particles, prepositions, and adverbs, and the other deals with the semantic and syntactic relations between the two components of the construction. Fraser (1965) defines particles as adverbial. He believes that phrasal verbs can be defined in syntactic terms alone. Following a transformational framework, Fraser lists eight transformations that distinguish a verb particle combination from a verb followed by a preposition. Phrasal verbs are defined in terms of the cohesion between the verb and the particle: if the combination is not cohesive then the element is a preposition. These eight tests which are summarized in Table 1.1, all support the conclusion that:

1. In *He looked up the information*, *look up* is a transitive verb + particle combination and *the information* is the direct object of the verb.

2. In *He looked at the table, at* is a preposition and *the table* its prepositional object.

The Transformation	Phrasal Verb	Prepositional Phrase
1. Relative &question	*Up what did he look ?	At what did he look ?
2. Manner insertion V-PART (modification)	*He looked quietly up the information.	He looked quietly at the information.
3. Object reduction (pronominalization)	*He looked up it. He looked it up.	He looked at it. *He looked it at.
4. PP conjunction	*He looked up the data and up the time of the event.	He looked at the table and at the chair.
5. Action Nominalization	His looking up of the information.	*His looking at the table.
6. Separation (particle movement)	He looked the information up.	*He looked the table at.
7. Manner insertion PART-NP (modification)	He looked up, without a reply, the information.	*He looked at, without a reply, the table.
8. Reduction & separation	He looked it up.	*He looked it at.

Table 1.1: Fraser's Tests[2]

According to Nilsen (1972), transformations 1, 2, and 4 above have the effect of keeping the particle with the following noun, but separating it from the preceding verb. For these transformations the verb+particle reading is ungrammatical, while the verb+preposition combination is allowed. Transformations 5 and 7 have the effect of keeping the particle with the preceding verb while separating it from the noun. This renders the prepositional

[2]PART is a part of speech label for particles.

combination ungrammatical, while the verb+particle combination remains acceptable. Transformations 4 and 6 look like a counter example to the cohesion argument, but Nilsen (1972:56) claims that these are used by Fraser to distinguish the separable phrasal verbs such as *call up (call her up)* from the non-separable ones such as *wait on* (**wait her on*).

There are several problems with Fraser's syntactic definition. Not all tests apply to all phrasal verbs. For example, according to test 2, modification is ruled out, but some modification is in fact possible, and manner insertion is acceptable in *John looked right up the information I asked for* according to Kayne(1985:127). In addition, the fact that transformations 4 and 6 apply to some phrasal verbs and not to others, raises the question of how conclusive these tests are. Nilsen challenges the eight tests proposed by Fraser, claiming that they are a failed attempt to classify particles and prepositions into non-overlapping categories. In Nilsen 1972 particles are defined as adverbial and a list of tests that differ from Fraser's is proposed and claimed to apply to all verb+adverb combinations. Nilsen emphasizes that exceptions do exist but he shows correlation between the different tests. In Nilsen's account there is no clear division between verb-particle association and verb preposition association, but rather a continuum. The continuum reflects the fact that not all tests apply although it seems that they will apply in the extreme cases (figurative combinations). Table 1.2 lists the tests that Nilsen uses to identify phrasal verbs.

Test	Example sentence
Separation	I called the boy up.
Separation & Object reduction.	I called them up.
Object reduction	I called on them.
Passive	The boys were called up (by someone).
Action	The calling up of the boys....
Gerundive	My calling up the boys...
Ambiguous phrase structure	The cat turned on + the radio The cat turned + on the radio
Adjectivalization	The called up boys...
Deletion of direct object	I called up
Deletion of P+NP	I ran out (of the house)

Table 1.2: Nilsen's Tests

Some of Nilsen's tests are identical to Fraser's. What is especially interesting is the ambiguous phrase structure test. Nilsen claims that in many cases a linear string can have both V+Prep +NP and V+PART+ NP structure and will be meaningful in both combinations. This constitutes a major problem for an NLP system where the input is a linear string and there is no prior information regarding the appropriate interpretation. The problem of ambiguity will be discussed further in chapter II.

The failure of syntactic tests to apply uniformly to phrasal verbs leads to a consideration of semantic alternatives. Fraser (1971) sets up three semantic categories of phrasal verbs: literal (*take up*), systematic (*stir up*), and figurative (*look up*). Co-occurrence

restrictions lead to two additional categories: Phrasal verbs that have the same co-occurrence constraints on the verb alone and on the combination and those which differ. He regards the particle as a special type of adverbial which in some cases should be part of the lexical entry of the verb. (A detailed account of Fraser's semantic analysis is given in chapter III.) Nilsen extends this notion and claims that some combinations are so idiomatic in nature that they require a separate listing in the lexicon.

Pelli (1972) uses a combination of syntactic and semantic tests to define phrasal verbs. He deals with the limited domain of stage directions and restricts his definition to apply best to his corpus. Pelli's results show that there is no test that will satisfy all requirements for phrasal verb selection. Two examples of tests which fail to account for different types of phrasal verbs are the syntactic passivization test and the semantic substitution test.

According to Pelli, verb + particle combinations that are interpreted figuratively are true phrasal verbs and should not passivize, while those with literal meanings should have passive counterparts. Consider the examples below where the a. combinations have literal meaning and the b. combinations have figurative meaning.

4a. He came down with his friends. (descended)

4b. He came down with a cold. (became ill)

5a. He finally got over the mountain. (crossed)

5b. He finally got over his cold .(recovered)

6a. He pulled up a heavy case. (lifted)

6b. He pulled up his car in front of the cafe. (stopped)

7a. Mary put up with John . (lodged)

7b. Mary put up with John . (tolerated)

If Pelli is correct, the literal passives below should be acceptable, and the figurative d. examples should be ruled out. But the passivization test applies inconsistently. There is no passive 4 because the verb *come* is intransitive. In addition, 5c is questionable and 7c unacceptable, while 6d and 7d are both acceptable.

4c. no passive

4d. no passive

5c. ?The mountain was gotten over. (crossed)

*5d. *His cold was gotten over.* (recovered)

6c. The heavy case was pulled up by him. (lifted)

6d. His car was pulled up in front of the cafe. (stopped)

*7c. *John was put up with by Mary.* (lodged)

7d. John was put up with by Mary. (tolerated)

Semantic substitution is the next test which Pelli considers. This test is used by several researchers to define phrasal verbs (Smith 1925, Palmer 1968, and Bolinger 1971). Many phrasal verbs can be replaced by one word verbs. For example *count out* can be replaced by *exclude* and *look into* by *investigate*. The problem is that it is often difficult and sometimes impossible to find a close paraphrase for some sentences. For example:

8. She went for him in a big way.

9. He hauled off and hit her.

In addition to the fact that substitution does not apply to all phrasal verb combinations it would be undesirable to generalize the substitution test as a general test for phrasal verbs because then combinations, like *eat dinner (dine)* and *make a mistake (err)* will have to be included. If Pelli is correct, then constructive substitution can help us identify combinations that are semantic units. For example if we replace *look up* (search) with *look at,* the meaning of the verb changes radically. Thus we know that *look up* is a semantic unit and *look at* is not. However, this test will not work in all cases since some forms do not allow contrast. For example:

10a. He cried out.

*10b. *He cried in.*

For these cases we can claim that the lack of contrastive evidence supports the claim that there are restrictions on the possible combinations of verb + particle.

As we see above, different types of semantic substitutions work with certain phrasal verb combinations but not with others, so the substitution test by itself cannot be used for phrasal verb selection.

Pelli found that for several reasons an overall test to extract phrasal verbs from his corpus was impossible to construct, first, because of the diversity in the syntactic structure of verb particle combinations, second, because of the possibility of figurative meaning in addition to the concrete meaning, and finally because of the fixed distribution (possible positions) of the particle. For his research, Pelli concentrates on three points: (1) the semantic distinction between literal and figurative usage should be preserved; (2) the phrasal verbs

chosen are the ones most commonly used[3]; and (3) the phrasal verbs will be able to undergo action nominalization. Pelli constructs a seven stage mechanism to screen the possible combinations in his corpus following the three guidelines above (and too many exceptions that are not relevant here since they do not follow any linguistic generalization). The significant point is that Pelli was forced to make these decisions because no test was able to account efficiently for the different types of phrasal verbs.

Bolinger (1971) claims that particles that form the most typical phrasal verbs are the ones that function sometimes as adverbs and sometimes as prepositions. He suggests mainly syntactic tests but also some semantic tests to distinguish verb+particle from verb+preposition combinations. Most of his tests were discussed earlier in the section on the work of Fraser and Nilsen before, but are listed in Table 1.3 below again. The new tests will be described in more detail.

[3]Pelli relies on the frequency of the phrasal verbs in his stage direction corpora to define 'common usage'.

Syntactic Test for Transitive V+PART	Example	Comments
Passive is possible	The information was looked up by the man. *The house is stood nearby the lake.	Applies only to distinguish PP not Adv.
Action Nominalization	His looking up of the information. *His looking at the table.	Applies only to distinguish PP not Adv.
Separation of V-PART	I took the weight off. (PART) *He sold regretfully the business.(Adv) *He stood the doorway in. (PP)	Can apply to some verb adverb combinations I walked over the hill. ? I walked the hill over. I walked the hill over from top to bottom.
Pronoun position	You're putting him on. *You're putting on him.	
No Particle modification (applies to both transitive and intransitive combinations).	The man drew the lucky number half way out. *The speaker drew his opinion only half way out.	The test separates independent adverbs from idiomatic phrasal verbs and can also reflect the degree of verb+particle closeness.
The definite NP test	see below	see below

Table 1.3: Bolinger's Tests

The definite noun phrase test is the most important test for Bolinger (adapted from Palmer 1968). It states that "A particle can precede a simple definite NP (proper name, or *the*+Noun) without taking it as an object" (Bolinger 1971:).

The examples below show how the test applies to verb-particle combinations (11, 12, 13),

but not to verb-adverb combinations (14, 15).

11. I am afraid to take on John in this contest.

12. You left out the captain.

13. Did you bring along the Joneses ?

14a. He finished up the report.

*14b. *He finished completely the report.*

15a. He pushed in the door.

*15b. *He pushed inward the door.*

The same order can be maintained, of course, in a preposition+noun combination, so there is a need to distinguish between the two. The distinction is best seen in a pronominalized structure. In the case of phrasal verb combinations the pronoun must be placed between the verb and the particle (16), while in the case of verb+preposition, the pronoun must follow the preposition (17).

16a. They brought on + the argument.

*16b. *They brought on it.*

16c. They brought it on.

17a. She told + on her friend.

17b. She told on him.

*17c. *She told him on.*

Bolinger regards the definite NP test as the most dependable criterion, and he considers every combination that passes this test as a phrasal verb. However, this test applies

to transitive combinations alone. Bolinger defines different semantic levels of phrasal verbs, but the definite NP test cannot distinguish between the different semantic levels that he proposes. (This issue is discussed further in chapter III.)

As stated earlier, Palmer (1968) acknowledges the need for a two level classification: the grammatical classification: the adverbial versus the prepositional nature of particles, and the semantic classification: whether the combination is idiomatic or not. He presents three criteria for the Adv-Prep grammatical distinction:

1. The adverb can appear both before and after the noun phrase in the following combination:

18a. He ran up a bill.

19a. He ran a bill up.

This is impossible with a preposition:

18b. He ran up the hill.

*19b. *He ran the hill up.*

2. When the object of the sentence is a pronoun, the order is fixed: the adverb must follow the pronoun as in, *He ran it up* (a bill), while the preposition must precede the pronoun, as in *He ran up it* (a hill).

Palmer uses the stress assignment criterion, which is also used by many others, to identify particles (Bolinger 1971, Pelli 1972, Fraser 1976, among others). Normally prepositions are never stressed, but particles usually are.[4]

18c. The hills I RAN up. (Prep)

19c. The bills I ran UP. (Adv)

---

[4]In this work the stressed words are capitalized.

Palmer indicates that we can find stressed prepositions to convey contrast as in:

20. The hills I ran UP are not the hills I ran DOWN.

The rule as formulated by Palmer is : "When in final position the preposition does not have nuclear stress (except in contrast) whereas the adverb has always nuclear stress" (p.)

Palmer discusses the behavior of four groups of possible candidates for phrasal verbs :

verb+adverb

verb+adverb+NP

verb+preposition

verb+preposition+NP

Palmer's basic assumptions that phrasal verbs have to have idiomatic meaning and that the particle is an adverbial lead to the classification shown in Table 1.4. The Table presents Palmer's four possible constructions for phrasal verbs of which are idiomatic in nature, and a combination of a verb and additional elements. They differ in their ability to passivize and the lexical nature of the combination. Only the first two columns are considered by Palmer as real phrasal verbs, while the others are prepositional verbs and phrasal prepositional phrases.

Phrasal Verbs		Prepositional verbs	Phrasal Prepositional Phrase
+Obj	-Obj		
V+ PART(Adv)	V+ PART(Adv)	V+PP	V+ PART(Adv) + PART(Prep)
+idiomatic	+idiomatic	+idiomatic	+idiomatic
+passive	+passive	+/- passive	+ passive
1 The bomb blew up 2. They carried on.	1. Put in an application. 2. put about a rumor.	Look after Look for Take to Go for	Put up with Do away with

Table 1.4: Palmer's Four Possible Candidates for Phrasal Verb Combinations

According to Palmer there are two types of phrasal verb combinations: phrasal verbs with an object and phrasal verbs without an object. These two types are considered in more detail below.

1. Phrasal verbs without object:

Palmer claims that particles are a certain type of adverbial, and so he tries to establish criteria to distinguish those combinations that are phrasal verbs from other verb+adverb combinations. Some verb+adverb combinations are idiomatic and some are not, as the examples below show:

21a. The bomb blew up. (idiomatic-explode)

21b. The flower pot blew down. (regular adverbial meaning)

More difficult cases of a three-way ambiguity where the same element can function as a particle, as a preposition, or as an adverb, can be also found:

22a. They carried on. (PART)

22 b. They carried on the business. (Prep)

23a. They turned over. (PART)

23b. They turn over the page. (Prep)

24a. He hung about. (PART)

24b. He hung about the place. (Adv)

One important structural distinction that can separate particle combinations from pure adverbial ones is the ability of an adverb and not of a particle to be preposed.

25a. He went away.

25b. Away he went.

26a. She goes out.

26b. Out she goes.

27a. He broke down.

*27b. *Down he broke.*

28a. He hung about.

*28b. *About he hung.*

This rule allows us to establish some combinations as verb + particle but not all; it is a good test that applies to some but not all phrasal verbs. For example, the sentence *He looked out* has an idiomatic meaning (pay attention) and a literal meaning (looked outward). Preposing is ungrammatical (**out he looked*), yet we should be allowed to treat *look out* as a

verb+adverb in its literal sense and as a phrasal verb in its idiomatic sense. The preposed form *out he looked,* however, can only be interpreted literally as *looked outward,* if it is allowed at all.

2. Phrasal verbs with object:

It is difficult to draw a line between the idiomatic and non-idiomatic senses of a combination that would allow us to distinguish between phrasal verbs and verb+adverb combinations. Consider some examples with the verb *put*:

29. Put back the clocks.

30. Put up a candidate.

31. Put off a meeting.

It is even more confusing because some combinations have more than one meaning, and others require the adverb to follow the noun:

32a. He ordered the man about.

*32b. *He ordered about the man.*

Neither Palmer nor the researchers previously discussed can find one satisfactory test that will apply to all phrasal verbs, and thus all are forced to make some arbitrary decisions regarding what they consider to be phrasal verbs. Palmer seems to rely more on the semantic idiomaticity of phrasal verbs as the basic criterion, but he cannot always determine whether the combination is idiomatic or not.

To conclude this section the important points of the discussion are summarized:

a. Most particles (e.g., *out, up, down, in,* etc.) can function sometimes as adverbials and sometimes as prepositions.

b. In the phrasal verb construction, the particle is a type of adverbial but not all adverbs can be particles.

c. Syntactic tests seem to be able to distinguish pure prepositions and pure adverbials from particles, but no test is conclusive and can account for all types of phrasal verbs.

d. The semantic notion of idiomaticity is a very important criterion. Phrasal verbs can be defined as idiomatic, but it is not always clear whether the combination is idiomatic or not.

e. Many combinations have both literal and figurative meanings resulting in an ambiguous structure.

f. It seems that a combination of semantic and syntactic tests can account for most of the cases of phrasal verbs. However, a definition for a restricted domain will be easier to arrive at.

g. A definition of phrasal verbs will be easier to state for a restricted domain because some semantic ambiguity can be resolved in context and because the number of combinations will be much smaller than in an unrestricted domain.

In the next section, I will present the definition of phrasal verbs used for the purpose of the research presented here.

1.3.3 A Definition of Phrasal Verbs for this Research

As explained in the previous section, there is no one definition of phrasal verbs which accounts for all aspects of their behavior. In addition, many different tests are suggested in the literature for identifying and classifying phrasal verbs, none of which apply to all types.

The purpose of this research is to capture as wide a range of occurrences of the construction as possible. For this reason, the definition of phrasal verbs will be stated as follows: A phrasal verb is a combination of a verb and a particle to which at least one of the following tests apply:

a. The verb and the particle can be separated.

33. I looked the information up.

b. A pronominalized object must appear between the verb and the particle.

34a. He looked it up.

*34b. *He looked up it.*

c. A phrasal verb can undergo an action nominalization.

35. His looking of the information up.

d. The phrasal verb can be passivized.

36. The boys were called up.

e. The phrasal verb can appear with or without an object.

37a. They carried on.

37b. They carried on the conversation.

f. The particle in the phrasal verb combination is usually stressed.

38. I looked it UP.

g. The phrasal verb has non-compositional semantics.

39. He called him up yesterday. (=to telephone)

h. The phrasal verb can be semantically substituted by a one word verb.

40. The bomb blew up. (=exploded)

i. There is strong distributional evidence that the verb and the particle are frequently associated (using statistical measurement over large corpora).

In previous treatments the verb and particle have been treated as two separate elements that may or may not combine. Much effort has been invested in trying to define the different types of particles and their semantic and syntactic effect on verbs. In this research we are concerned with the phrasal verb as a whole, and great emphasis will be put on the relationship between the verb and the particle. The assumption is that the classification of phrasal verbs will have to take into consideration this relationship. This should be reflected in the representation of the construction in the lexicon. Even in cases where the verb and the particle each have their own entries, there must be some reference as to which verbs and particles can combine and the nature of the combination in terms of the verb+particle relationship.

The next chapter is devoted to describing the problem of ambiguity in general and phrasal verb ambiguity in particular in reference to natural language processing systems. The purpose is to show how common ambiguity is in all levels of processing and how one mistake can lead to many more mistakes as they are carried from one processing level to another.

CHAPTER II: THE PROBLEM OF PHRASAL VERB AMBIGUITY

2.1 INTRODUCTION

The basic claim of this chapter is that phrasal verb ambiguity is a challenge for any NLP system, whether rule based or statistical. This chapter will explore the problem of representing and processing phrasal verb constructions in various NLP systems, emphasizing the effect of mishandling the verb+particle combination.

Sentences that pose no difficulty to humans, due to their extensive world knowledge, may cause ambiguity problems during natural language processing by a machine.

Below are some examples of the most common types of ambiguities:

A. **Lexical ambiguity:** also called categorical or part of speech (POS) ambiguity. It is reflected in the different POS or tags assigned to a word. If taken in isolation, most words in English text are categorically ambiguous, i.e., they can belong to more than one part of speech.

For example: *to* can be the complementizer of an infinitive or a preposition, as seen in the following sentences:

1a. I would like to visit Paris. (infinitive)

1 b. I walked all the way to the office. (preposition)

B. **Phonological ambiguity**: One written form is pronounced differently yielding two words, for example *wind* can be /wind/ or /waind/.

C. **Structural ambiguity**: A sentence may have more than one structural assignment. For example in the following sentence the prepositional phrase can have two possible attachments:

2a. I saw the man with the telescope.

2b. I saw the man [$_{pp}$with the telescope] (=I used a telescope to see the man.)

2c. I saw [the man with the telescope] (In answer to which man did YOU see?)

In 2b the PP is the complement of the verb phrase and the interpretation conveyed is of manner modification (i.e., *how did I see*).

In 2c the PP modifies the NP itself to convey the meaning *Which man did I see?*

D. **Word Sense ambiguity**: A word might have more than one meaning or be associated with different senses depending on the context, for example:

3a. The BANK of the river is rocky.

b. I closed my account at the BANK today.

E. **Scope ambiguity:** The scope of a quantifier is not always clear, for example:

4. John found a defect in every car with over 500 miles. (same defect?)

5. Every one in the room spoke two languages. (same two languages?)

A distinction is often made between local and global ambiguity. Local ambiguity means an item can be ambiguous when initially encountered, but this ambiguity can be

resolved once the rest of the sentence is viewed. Lexical ambiguity is one example of local ambiguity. Since in English most words can be assigned more than one POS, the result is a large amount of local ambiguity, as we see in the following examples:

6. The management requests(V/N) control(V/Adj) information.

Global ambiguity is independent of the words and is inherent in the sentence in general. For example, in structural ambiguity two possible structures can be assigned to a sentence although the words and their POS are identical. Global ambiguity usually occurs in regard to prepositional phrase attachment, coordinate structures, and compound structures, as illustrated below:

7a. Good boys and girls go to heaven.

7b. Good [boys and girls] go to heaven.

7c. [Good boys] and girls go to heaven.

8a. Joe read the book on the floor.

8b. Joe read the book [on the floor].

8c. Joe read [the book on the floor].

A sentence might have three types of ambiguity. For example:

9a. *I saw her duck.*

This sentence exemplifies semantic, lexical, and structural ambiguity. The semantic ambiguity is reflected in the fact that *saw* can mean past tense of *to see* or present tense of *to saw*. *Duck* is lexically ambiguous between noun and verb. And the structural ambiguity is reflected in two possible attachments of *her* as possessive pronoun referring to *duck* (9b) or as a personal pronoun acting as the object of *saw* (9c):

9b. I saw [her duck].

9c. [I saw her] duck.

There are more types of ambiguity, but the important point to be made here is that ambiguity arises in all areas of computational linguistics and every NLP system has to deal with ambiguity problems in order to process unrestricted language.

Two different approaches were taken to handle Lexical Disambiguation. Natural language processing and artificial intelligence (AI) in general, have focused on building rule-based systems with carefully handcrafted rules and domains of knowledge. In recent years there is been a growth in the use of probabilistic models for natural language processing. Taggers based on statistical evaluation have been developed in response to the difficulties of traditional rule-based systems to handle large-scale applications involving unrestricted text. Taggers (e.g., Beale 1988, Church 1988) offered an alternative approach to the handcrafted rules and intractable amount of work put in by linguists. The rationale underlying this approach is that what is probable is good enough to cover most of the language used and that rule-based systems involve too much work in trying to account for, uncommon constructions.

The applications of statistical language models pertain to many NLP areas, including: speech recognition and synthesis, parsing, information retrieval, and machine translation. In all of these areas ambiguity resolution is an important problem. The next section will provide some background to the historical development of statistical models for NLP.

2.1.1 Statistical Models in NLP

Liberman (1991) traces the "trends towards statistical models in NLP" and describes their historical development.in the past four decades. Statistical models in natural language flowered in the 1950s and early 1960s; they were then abandoned in the 1970s and most of the 1980s, but have attracted growing attention since the late 1980s.

In the 1950s empiricism and statistical models were at their peak. But early stochastic models were extremely naive, simple, and obvious and were highly criticized by rival approaches. Some rejected the models on logical grounds, claiming they did not distinguish among the conceptually different sorts of information in syntactic, semantic, and pragmatic constraints. In *Syntactic Structures* (1957) Chomsky rejects the use of frequency analysis for natural language. He presents three basic assumptions about knowledge of language that will pose a problem for statistical models: The first assumption is that it is sufficient to assume partial knowledge of sentences and non-sentences. Some sentences clearly belong or dont belong to the language. Secondly, each grammar is related to the corpus of sentences in the language. Thirdly, it is not necessary to know all the existing sentences in order to build a grammar. On the basis of a typical set of sentences from language L we can build a grammar adequate for language L in particular and language in general.

According to Chomsky, these assumptions will pose a problem to the statistical model since it does not contain a grammar, and as a result the model has no knowledge of the language structure or its grammar . It is merely a collection of tables of frequencies and probabilities. Thus, it is not sufficient for such a system to encounter few sentences in order to draw conclusions about the structure of other sentences. Knowledge is gained in such a

system by direct exposure to data. A statistical model is not able to distinguish ungrammatical from grammatical sentences, nor is it able to distinguish levels of grammaticality. Chomsky cites as an example the grammatically well-formed but semantically ill-formed sentences:

10. Colorless green ideas sleep furiously.

Chomsky concludes his discussion in *Syntactic Structures* by saying that Probabilistic models give no particular insight into some of the basic problems of syntactic structure. Yet he agrees that statistical models can be useful once the syntactic structure and the grammar of the language have been determined. This formulation has become known as "Chomsky's Hierarchy," and has attracted much attention from researchers.

Interest in information theory also declined due the high cost of computer resources and their low computational power at the time. Data were limited and there were no large corpora available, so that results were neither powerful nor representative. Liberman (1991) suggests that it was an "anti-empiricist", "anti-numerical, pro-symbolic" era, and "counting" was not highly regarded. For all these reasons, interest in stochastic models and in corpus-based linguistics declined drastically.

All this started to change in the beginning of the 1980s. The speech recognition community led the way. Researchers like Jim Baker and Fred Jelinek at IBM used stochastic methods based on Shannon's noisy channel model for speech recognition. New mathematical techniques not known in the 1950s, such as re-estimation methods for Hidden Markov Models (HMM) and Stochastic Context Free Grammars (SCFG), were now available as were techniques for inducing stepwise optimal decision trees, and improved estimation procedures

for dealing with the sparse data characteristics of linguistic distribution. Another reason for this empirical renaissance was the increasing availability of massive quantities of data, from the Brown Corpus (Francis & Kucera, 1982) , a one million word corpus to today's text collections which run into hundreds of millions of words. Large corpora allows data to be examined on a much larger scale, and provide coverage of a greater variety of linguistic phenomena, which in turn allow for better training of the analyzer on a wider range of different types of text. All of this is possible because of the development of powerful computer technology and its falling cost in recent years.

2.2 THE IMPORTANCE OF RESOLVING PHRASAL VERB AMBIGUITY IN AN NLP SYSTEM

The claim among researchers advocating the use of statistics for NLP (e.g., Marcus et al., 1992) is that taggers are routinely correct about 95% of the time. The 5% error rate is not perceived as a problem mainly because human taggers disagree or make mistakes at approximately the same rate. On the other hand, even a 5% error rate can cause a much higher rate of mistakes later in processing, particularly if the mistake falls on a key element that is crucial to the correct analysis of the whole sentence. One example is the phrasal verb construction (e.g., *gun down, back off*). An error in tagging this two-element sequence will cause the analysis of the entire sentence to be faulty. An analysis of the errors made by the stochastic tagger PARTS (Church, 1988) reveals that phrasal verbs do indeed constitute a problem for the model.

The following sections explain the nature of the lexical, syntactic, and semantic ambiguity of phrasal verbs and the consequences of ignoring this ambiguity. The first section will discuss tagging issues and will present the problem of representing the phrasal verb and identifying it, whether the work is done by human taggers as in the Brown Corpus (Francis and Kucera 1982) and the PennTreebank (Marcus, et al. 1992) projects, or by statistical approximations as in PARTS. The mistagging of phrasal verbs can result in the assignment of incorrect structure and also in misinterpretation. Furthermore, other operations like speech synthesis which rely on tagging for stress prediction, can be affected. Section 2.2.1.2 will discuss such an example in detail.

Sections 2.2.2 and 2.2.3 will discuss the difficulty in correctly representing the syntactic

structure and the semantic nature of phrasal verbs in the computational lexicon, mainly due to the variety of behaviors of the construction. The relationship between the syntactic and semantic nature of phrasal verbs must be explored in order to correctly represent them in the lexicon.

2.2.1 Tagging Issues

In this section we are interested in the task of part of speech (POS) tagging and how to solve the ambiguity resulting from an element having more then one POS, as described in section 2.1. While human taggers rely on both linguistic knowledge and intuition in assigning POS tags, statistical part of speech taggers use the probabilities of the independent words and probabilities in context to account for the most probable tag. These taggers are usually initially trained on manually tagged samples.

Statistical taggers are commonly used to preprocess natural language. Operations like parsing, information retrieval, machine translation, and so on, are facilitated by having as input a text tagged with a part of speech label for each lexical item. In order to be useful, a tagger must be accurate as well as efficient. In the next section one such tagger PARTS is described.

2.2.1.1 The Stochastic Tagger PARTS

This section describes the tagger PARTS for the purpose of discussing the problem of lexical ambiguity. The algorithm used in PARTS is described, and the problems of identifying and classifying phrasal verbs are discussed. In Chapter IV, some solutions to

these problems are suggested.

Church (1988) advocates using frequency information to resolve lexical ambiguity. He claims that dictionaries and grammar-based systems tend to focus on what is possible, not on what is likely, thus wasting computational power perusing paths of very low probability.

PARTS is a stochastic tagger similar to other statistical taggers, such as De Ross (1988) and Meeter, Schwartz, and Weischedel (1991). These taggers are based on Shannon's noisy channel model (Shannon, 1948). This model is used in speech recognition applications to determine the input words given the noisy output. By analogy with the noisy channel formulation, the input is the tags and the output is the words, as in:

11a. tags----->noisy channel---->words

The role of the tagger is to determine the tags from the given words, which is the opposite direction of the noisy channel model. For the tagger's operation input is the words and the task of the program is to output tags.

11b. tags<------tagger<--------words

These programs input a sequence of words and output a sequence of POS tags. For example: (Church et al., 1993:6)

12a. Input words: *The chair will table the motion*
| | | | | | |
Output Tags: *art noun modal noun art noun*

The PARTS program uses dynamic programming to find assignments of POS to words. The tagger makes use of both lexical and contextual probabilities in a trigram model (three window buffer) to arrive at the most probable tagging assignment. The input to the statistical approximations are the inflected forms of the words (rather than the stems). Church

claims that the probability of the whole are not easy to derive from the parts and that in fact the relationship between the probability distribution of the base form and the inflected form is much more complicated than one might have expected[5].

In theory, the proper algorithm should include the probability distributions Pr(P) and Pr(W|P). Unfortunately, these probability distributions are very complex: Pr(W|P) is a table giving every pair W and P of the same length a number between 0 and 1, which is the probability that a sequence of words chosen at random from English text and found to have the part-of-speech sequence P will turn out to be the word sequence W. Changing even a single word or part of speech in a long sequence may change this number by many orders of magnitude.

Church et al. (1993) claim experience has shown that high tagging accuracy can be achieved in practice using the following simple approximations where P_i is the i^{th} part of speech in the sequence P, and W is the i^{th} word in W.

Contextual Probability: the probability of observing POS P given k adjacent parts of speech (in a trigram k=2).

$$13.\ Pr(P_1,P_2,....P_N) \approx \prod_{i=1}^{N} Pr\,(P_i|P_{i-2}\,P_{i-1})$$

Lexical Probability: the probability of observing POS P given word W:

$$14.\ Pr(W_1, W_2,...W_N|P_1, P_2,....P_N) \approx \prod_{i=1}^{N} Pr(W_i|P_i)$$

Pr(P) is replaced by a trigram approximation and P(W|P) by an approximation in which each word depends only on its own part of speech.

The parameters of this model, the lexical probabilities, and the contextual

---

[5]K.Church, August 1994, Personal communication.

probabilities are generally estimated by computing various statistics over large bodies of text. The search for the appropriate POS is done the following way: the search enumerates all possible assignments of POS to input words are enumerated, producing set of possible assignments. Each possibility is then scored by the product of the lexical probabilities and the contextual probabilities and the best sequence is selected. In some cases it is possible to know that certain sequences cannot possibly compete with others and should therefore be abandoned. This shortens the process.

Word	Parts of Speech	
I	pronoun 5837[6]	proper noun 1
see	verb 771	interjection 1
a	article 23013	in (French) 6
bird	noun 26	

Table 2.1 Sample of Word Frequencies in PARTS[7]

Church (1988) claims that PARTS performs correctly 95% of the time, yet there are clearly some problems that might result in a higher error rate for certain constructions. One problem is that in POS disambiguation we need to estimate how often each word appears

[6]This means that the probability that *I* is a *pronoun* is estimated as the frequency 5837/5838 (true cases out of all cases).

[7]Table is taken from Church 1988:139.

with each POS. Even if we look at a large corpus, there will always be words that appear only a few times, so that it is impossible to estimate the probability of their being one POS or another without more information. Church suggests consulting a dictionary in these cases.

Another problem which is more crucial for the analysis of phrasal verbs is that the basic assumption underlying the stochastic process is the notion of independence. The statistical approximations apply to words, and in the tagger, words are separated by spaces. As a result the components of the phrasal verb are treated as two individual words each having its own lexical probability. An error analysis performed on the tagging results of PARTS shows an interesting pattern. A phrasal verb such as *sum up* will be tagged by PARTS as noun+preposition instead of verb+particle. This error influences the tagging of other words in the string as well. One typical error is found in infinitive constructions, where a phrase like *to gun down* is tagged as INTO NOUN IN (a preposition followed by a noun followed by another preposition). The reason for this error is that words like *gun, back,* and *sum,* by themselves, have a very high probability of being nouns as opposed to verbs. However, when they are followed by a particle they are usually verbs, and in the infinitive construction they are always verbs. The error appears to follow from the algorithm of the stochastic program itself. As we said above, in the trigram model the probability of each word is calculated by taking into consideration two elements: the lexical probability (probability of a word bearing a certain tag), and the contextual probability (probability of a word bearing a certain tag given two previous elements). As a result, if a word has a very high probability of being a noun (*gun* is a noun in 99 out of 102 occurrences in the Brown Corpus) it will not only influence, but will actually override the contextual probability, a

different part of speech assignment. Thus that the tagger will perform poorly on phrasal verbs in those cases where the ambiguous word is occurs more frequently as a non-verbal element, but will have fewer problems handling this construction when the ambiguous word is a verb in the vast majority of the cases. The tagger's POS assignment for the task of labeling phrasal verbs should be analyzed to determine whether the model should be modified to take into consideration the dependency between the verb and the particle in order to optimize performance. Shaked (1993) describes a test which shows that the accuracy level of PARTS for tagging phrasal verbs is 89% although Church (1988) claims that PARTS overall performance is 95%-99% correct. If this is the case, then for the task of labeling phrasal verbs PARTS' performance is below average. In addition, as mentioned earlier, a mistake in identifying and labeling the phrasal verb, especially where it is the main verb of the sentence might cause the analysis of the whole string to fail in applications that take tagged text as an input, thus resulting in a much higher error rate.

The next section deals with an example of this in the domain of speech synthesis. Models for predicting prosodic structure make use of statistical taggers such as PARTS, and thus any mistake in POS assignment is carried on to the next processor or application.

2.2.1.2 Tagging and Speech Synthesis

In English, lexically stressed syllables tend to alternate with unstressed syllables to create a rhythm. Within words and phrases some syllables are more prominent than others. When a word is spoken as part of a longer utterance, intonation patterns will make some

lexically stressed syllables more prominent by placing a pitch accent and these accents will indicate phrases within an utterance. A syllable cannot have a pitch accent unless it also has lexical stress.

Numerous factors affect accent placement, including syntax, discourse structure, and rhythm. Both Ross (1992) and Hirschberg (1993) show POS assignment is by far the most important feature for determining accent location. The study conducted by Ross et al. (1992) examined the factors that can influence whether or not a word is accented., these include the class of the word, its adjacent words, and its position in a large prosodic constituent. This approach is different from most traditional rule-based methods of accent prediction. The problem with this type of approach, according to Ross, is that it allows only two types of pitch accent (accent and nuclear accent) and does not separate the different types of pitch accents (such as high, low, high-low, low-high, etc.)

High quality speech synthesis is needed to make text-to-speech technology useful in more applications. Prosody , the super-segmental level of speech which also supplies semantic information, is one of the aspects of speech synthesis needing improvement.

Ross (1992) presents an empirical study of the factors that affect accent placement in American English. His research aims at supporting the development of a computational model for predicting pitch accent location in speech synthesis. This automatically trainable model could be then incorporated into an existing system, and will include two modules: one to predict abstract prosodic labels from text, and one to generate fundamental frequency and energy contours from abstract labels. The research draws on both linguistic theory and on statistical modeling to provide a mechanism for automatically generating the model

parameters. Specifically, decision trees are used to predict the abstract labels and a dynamic system model with parameters at three time scale is used to generate the fundamental frequency and energy contours. In this account I will be concentrating on the first task only since it is of interest to this research.

What is most relevant to this researche is the POS factor and how the tagging of phrasal verbs might influence prosody. In Ross' system, the first division into word classes is between content words and function words. Traditional text-to-speech systems have accented content words and de-accented function words. By definition, function words have little semantic content of their own; they are a closed class and serve primarily to mark grammatical relationships in the sentence. Ross' function words category includs: determiner, auxiliary verb, negative, preposition, pronoun, possessive pronoun, qualifier quantifier, participle, conjunction, modal, and wh-word. All other words belong to the open class of content words. It is not clear where particles belong in this scheme.

The corpus for Ross' experiment is a set of recorded FM radio news broadcasts, spoken by two female announcers, which contain 7880 words and 3861 pitch accents. The text was annotated with part of speech tags obtained automatically using a statistical tagger developed by BBN, and PennTreebank labels. Ross' goal was to build a model that would predict which part of speech was likely to be accented.

His first observation was that, in spite of the generalization about de-accenting function words, some function words are accented. He found that 39% of the words in the corpus were function words, of which 11% had pitch accent. Most of this 11% were negatives *(not, no, never, etc.)* and quantifiers (*all, every, many, etc.*). This prompted Ross

to further divide function words into two subclasse: those that are often accented and those that are rarely accented.

POS is indeed only one factor in the prediction model, yet Ross et al. (1992) claim that it is the most important factor after lexical stress. How important this information is for the prosodic labeling of phrasal verbs will be the concern of the next part of this section.

We would like to evaluate the importance of correct POS tagging in a prediction model for the prosody of phrasal verbs (Ross 1992). In this section I present a set of examples(provided by Ken Ross), which includes (a) four example sentences where particles were incorrectly tagged as prepositions, resulting in synthesized speech with clear prosodic errors caused by his prediction model, and (b) variations of these sentences with correct particle labeling. In the first two cases the particle tag was correctly assigned by the tagger (POST), and in the second two cases it was incorrectly identified as a preposition. (RP= particle, IN=preposition).

15a. But a new application of the technology is about to be tried out in Massachusetts to ease crowded jail conditions.

15b. a/DT new/JJ application/NN of/IN the/DT technology/NN is/VBZ about/RB to/TO be/VB tried/VBN out/RP in/IN Massachusetts/NP to/TO ease/VB crowded/JJ jail/NN conditions/NNS.

16a. Next week some inma tes released early from the Hampton County jail in Springfield will be wearing a wristband that hooks up with a special jack on their home phones.

16b. Next/JJ week/NN some/DT inmates/NNS released/VBD early/RB from/IN the/DT

Hampton/NP County/NP jail/NN in/IN Springfield/NP will/MD be/VB wearing/VBG a/DT wristband/NN that/WDT hooks/VBZ up/RP with/IN a/DT special/JJ jack/NN on/IN their/PP$ home/NN phones/NNS.

17a. Whenever a computer randomly calls them from jail, the former prisoner plugs in to let corrections officials know they're in the right place at the right time.

17b. a/DT computer/NN randomly/RB calls/VBZ them/PP from/IN jail/NN the/DT former/JJ prisoner/NN plugs/NNS in/IN to/TO let/VB corrections/NNS officials/NNS now /BP they're/PP in/IN the/DT right/JJ place/NN at/IN the/DT right/JJ time/NN.

In example 17b we notice that the tagger not only misidentified the particle but missed the verb and classified it as NNS (plural noun). These two mistakes are probably linked as a result of the statistical approximation, the same phenomenon we witnessed in PARTS.

18a. Those on early release must check in with corrections officials fifty times a week according to Ash, who says about half the contacts for a select group will now be made by the computerized phone calls.

18b. Those/DT on/IN early/JJ release/NN must/MD check/VB in/IN with/IN corrections/NNS officials/NNS fifty/CD times/NNS a/DT week/NN according/VBG to/TO Ash/NP who/WP says/VBZ about/RB half/PDT the/DT contacts/NNS for/IN a/DT select JJ group/NN will/MD now/RB be/VB made/VBN by/IN the/DT computerized/JJ phone/NN

calls/NNS.

As a result of the incorrect POS tagging the prediction model, which uses the data from the tagger, will continue to carry the mistake to the prosody assignment phase as well. This will lead to an incorrect intonation prediction, and the synthesis will be affected as well.

The following examples show the predication results for each sentence as calculated by Ross model. We see the difference in accent prediction depending on which the tag is assigned to the particle. These predictions are represented in the prosody column, which indicates the type of accent predicted to be assigned to the current word[8]. If the tagger incorrectly labeles the particle as a preposition the prediction is that the word will not be accented and accent falls on the preceding verb. In the correct version when the particle tag is assigned, the prediction model not only accents the particle but predicts a phrasal boundary immediately after the particle and deaccents the preceding verb. Although the intonational effect of the error is local to the verb and the particle, the interpretation of the whole sentence is adversely affected when the phrasal verb is not identified and therefore accented incorrectly.

The model uses the TOBI, a 3 tier labeling system which uses break indices to represent different intonational levels 1-6.
0- criticized word pair
1- no intonational mark
2- not a real boundary but slightly accented.
3- intermediate phrase boundary
4- intonational phrase boundary
5- a phrasal boundy shorter than a sentence break (equal to semicolon)
6- sentence break

Correct Particle Tagging			Incorrect Particle Tagging		
word	**pos**	**prosody**	**word**	**pos**	**prosody**
Sentence #15					
But	CC	(s) 1			
a	DT	(s) 1			
new	JJ	(P) 1			
application	NN	(P s P L) 3			
of	IN	(s) 1			
the	DT	(s) 1			
technology	NN	(s P s L) 3			
is	VBZ	(s) 1			
about	RB	(s P) 1			
to	TO	(s) 1			
be	VB	(s) 1			
tried	VBN	(P) 1	tried	VBN	(P-L) 3
out	RP	(P-L) 3	out	IN	(s) 1
in	IN	(s) 1			
Massachusetts	NP	(s s P L) 3			
to	TO	(s) 1			
ease	VB	(P) 1			
crowded	JJ	(P s) 1			
jail	NN	(P) 1			

conditions	NNS	(s P CR) 6

Sentence #16

Next	JJ	(P) 1
week	NN	(P-!H) 3
some	DT	(s) 1
inmates	NNS	(P s) 1
released	VBD	(s P) 1
early	RB	(P L) 3
from	IN	(s) 1
the	DT	(s) 1
Hampton	NP	(P s) 1
County	NP	(P s) 1
jail	NN	(P-L) 3
in	IN	(s) 1
Springfield	NP	(P !H) 3
will	MD	(s) 1
be	VB	(s) 1
wearing	VBG	(P L) 3
a	DT	(s) 1
wristband	NN	(P !H) 3
that	WDT	(s) 1

hooks	VBZ	(P-H) 3	hooks	VBZ	(P) 1
up	RP	(P-!H) 3	up	IN	(s) 1
with	IN	(s) 1			
a	DT	(s) 1			
special	JJ	(P s) 1			
jack	NN	(P-L) 3			
on	IN	(s) 1			
their	PP$	(s) 1			
home	NN	(P) 1			
phones	NNS	(P-FF) 6			

Sentence #17

Whenever	WRB	(s s s) 1
a	DT	(s) 1
computer	NN	(s P s) 1
randomly	RB	(P s s) 1
calls	VBZ	(P) 1
them	PP	(s) 1
from	IN	(s) 1
jail	NN	(P-CF) 4
the	DT	(s) 1

former	JJ	(P s) 1
prisoner	NN	(P s s) 1
plugs	NNS	(P-L) 3
in	RP	(P-H) 3
to	TO	(s) 1
let	VB	(P) 1
corrections	NNS	(s P s) 1
officials	NNS	(s P s) 1
know	VBP	(P) 1
they're	PP	(s) 1
in	IN	(s) 1
the	DT	(s) 1
right	JJ	(P) 1
place	NN	(P-L) 3
at	IN	(s) 1
the	DT	(s) 1
right	JJ	(P) 1
time	NN	(P-FF) 6

plugs	NNS	(P-L) 3
in	IN	(s) 1

Sentence #18

Those	DT	(s) 1			
on	IN	(s) 1			
early	JJ	(P s) 1			
release	NN	(s P-L) 3			
must	MD	(s) 1			
check	VB	(P) 1	check	VB	(P-H) 3
in	RP	(P-!H) 3	in	IN	(s) 1
with	IN	(s) 1			
corrections	NNS	(s P s) 1			
officials	NNS	(s P s) 1			
fifty	CD	(P s) 1			
times	NNS	(P-L) 3			
a	DT	(s) 1			
week	NN	(P) 1			
according	VBG	(s P s) 1			
to	TO	(s) 1			
Ash	NP	(P-CF) 4			
who	WP	(s) 1			
says	VBZ	(P-H) 3			
about	RB	(s P) 1			
half	PDT	(P) 1			
the	DT	(s) 1			

contacts	NNS	(P L) 3
for	IN	(s) 1
a	DT	(s) 1
select	JJ	(s P) 1
group	NN	(P-L) 3
will	MD	(s) 1
now	RB	(P-H) 3
be	VB	(s) 1
made	VBN	(P-H) 3
by	IN	(s) 1
the	DT	(s) 1
computerized	JJ	(s P s s) 1
phone	NN	(P) 1
calls	NNS	(P-FF) 6

Hirschberg (1993) deals with very similar issues. The goal of his study is to determine what sort of success can be achieved in predicting pitch accent in natural speech simply by using information that can be obtained automatically from text using current computational techniques, and which types of information are most useful.

Hirschberg claims that POS information is the most employed predictor of pitch accent. Most text-to-speech systems assign accent based on a simple distinction between function words and content words, as we saw in Ross et al. (1922). Hirschberg points out that

items that are ambiguous in respect to the function/content distinction are rarely disambiguated, since text-to-speech parsers and taggers tend to be extremely simple. Such an example is the case of preposition/particle ambiguity, where prepositions are commonly deaccented and particles are commonly accented.

Hirschberg deals with ambiguity problems by adding to the tagger PARTS special information about problematic constructions. So in the case of preposition/particle ambiguity her model incorporates a look-up table for verb particle constructions (special list).For those phrasal verbs identified by the rule, the verb is deaccented and the particle gets the accent instead. In summary, Hirschberg uses hand-crafted rules to improve pitch accent prediction. If disambiguation is dealt with in the tagging phase, there will nott be any need for hand-crafted rules in the prediction model. On the other hand, this is an example of how hand-crafted rules can improve the results of statistical approximations.

2.2.2.3 Tagging in other systems

The status of the phrasal verb constituent is still an unresolved issue in linguistic theory, so it is hardly surprizing that it also poses a problem for natural language systems. It is interesting to see how the phrasal verb construction is handled there.

The first tagging project to consider is the Brown Corpus (Francis & Kucera 1982). The motivation behind creating the tagged Brown Corpus was the absence of a consistent system of overt marking of word class constituency and the fact that words may easily belong to different word classes without any morphological change. In tagging the Brown Corpus

there were two main phases. The first procedure was the automatic grammatical tagging, which resulted in the accurate tagging of 77% of the corpus. The remaining 23% was completed manually over a period of nearly ten years. The tag set used included 87 tags of five kinds: major part of speech classes (nouns, verbs, etc.); function words (determiners, prepositions); certain important, individual words (*not, to, have,* etc.); punctuation marks; and four designator tags (for special forms of discourse). The Brown Corpus is important because is provides the basis for a great deal of later work. Taggers such as PARTS were trained on the Brown Corpus and the tag set has served well as a model for other projects, such as the PennTreebank.

The annotators of the Brown Corpus treated the verb+particle sequence as two separate elements. "In the case of a two part verb the attempt was made to distinguish between adverbs (Hold out your hand) and adverbial particles (Hold out for more money). It was found, however, that this necessitated a large number of arbitrary decisions, which might confuse or mislead those using the corpus. It was decided instead to consider this a semantic rather than a taxonomic problem and to give the tag RP (for either adverb or particle) to the ten words: about, across, down, in, off, on, out, over, through, up. The exception is when these words function as a preposition which in this case they are tagged as normal preposition." (Francis & Kucera 1982:13).

The way the problem of phrasal verbs was handled here was to consider the phrasal verb as a semantic rather than a syntactic unit, and to assume that a phrasal verb can contain a verb followed by either an adverb or a particle, thus delegating the responsibility of disambiguation to the semantic component rather than dealing with the structural issue.

In the The PennTreebank project (Marcus et al. 1992) the corpora were annotated by human taggers. The taggers handled verb+particle combinations by considering the verb

and the particle as one unit and not annotating the particle at all, thus assuming it is part of the verb as in [$_{V}$hand over]. Here we see that the decision was made to treat the sequence as a phrase, if not for theoretical reasons then for practical reasons.

2.2.2 Syntactic Structure

2.2.2.1 ROBIE (Milne 1983)

ROBIE is a grammar-based parser which handles phrasal verbs in the following way: For each verb there is a list of potential particles such that if one occurs with the verb it must be used as a particle. To assure that a high priority rule operates in the packet SS-VP with a pattern: *19. [particle],sem ----> chk(particle).*

This rule will apply if the semantic check will go over the obligatory particles associated with the semantics of the main verb. If the particle is not on the list, the rule will not apply and the particle will be left in the buffer to be picked by the preposition (PP) rules. If the particle is not picked up by the PP rules it will stay in the buffer and later will be attached to the verb phrase as a particle. The same rule will deal with particle movement.

Milne's analysis has some problems. The first problem is that the semantic check is based only on a list of verb+particle pairs. To be correct, the list should include the optional and obligatory prepositions that the verb can take as well, and then make a more clever decision. This creates the second problem, which is if the particle is not on the list and the priority rule is not matched, the particle will remain in the buffer and will be picked up by the PP rules. If the particle is followed by an NP starter (like a Determiner, for example) as in *Please pick up the mail*, it will be analyzed as a preposition. Since this is a very frequent

combination the error rate will be high. The third problem is that no attempt is made to ensure the unacceptability of ungrammatical preposition/particle ambiguity as in :

20a. He sat on the chair.

*20b. *He sat the chair on.*

Furthermore, some preposition/particle ambiguity is global and will only be resolved based on a non-syntactic choice when the next noun phrase is parsed semantically. As in:

21. He looked up the street.

a. gazed.

b. searched for.

Milne acknowledges these flaws and in handling phrasal verbs in his system, but claims that in order to parse verb+particle combinations correctly, better verb subcategorization frames are needed, as well as more detailed selectional restrictions on the verbs.

Milne's main criticism of the mishandling of verb+particle combinations is that the focus of linguistic theory is mostly on the type of particle allowed in the combination. Milne claims that if we are to treat the sequence as one unit, than the relationship between the verb and the particle should be further explored from the point of view of the selectional restrictions and subcategorization frames of the verb.

The following questions are relevant to this issue. Can we define common restrictions for certain verbs and not others? Do verbs with similar selectional restrictions have other features in common, that is, do they constitute natural classes ? Does the current

classification of particles matches by the classification of verbs. Is this a bi-directional selection, meaning that the restrictions on the possible combinations is determined by both the verb and the particle, and not by only one of them. The linguistic literature I researched touched on these questions briefly and but did not give conclusive answers that would enable us to determine the relationship between the two elements of the phrasal verb combination.

2.2.2.2 COMLEX: a Syntactic Lexicon

The purpose of this section is to show how phrasal verbs are treated in the syntactically-based lexicon, COMLEX Syntax (Macleod and Grishman 1994). The structure and components of the dictionary will be explained in detail, concentrating on the representation of phrasal verbs. This will be followed by a discussion of the advantages and disadvantages of this project's approach. The main criticism of this dictionary is that being based on syntax alone it lacks the ability to fully treat all types of phrasal verbs with accuracy.

COMPLEX is a project whose purpose is to develop a moderately broad coverage English dictionary containing syntactic features needed for natural language analysis. The assumption is that this dictionary should be coupled with pronunciation and semantic components. COMLEX Syntax (hereafter COMLEX) in its current state is designed to have lexical entries only for nouns, verbs, and adjectives.

A dictionary entry takes the form of a parenthesized, nested, feature-value structure. The POS type of the entry is indicated at the top, and if one word has more than one POS it will have several entries. Each entry includes the following four components:

1. ORTH: An orthography field which contains the spelling of the word.

2. SUBC: Subcategorization frames (required for verbs, optional for nouns and adjectives). The SUBC field is further divided into the following components:

a. TYPE: Major constituent type (POS) of the entry. The system uses 13 POS tags: adjective, adverb, advpart (adverbial particle), aux (auxiliary), cconj (coordinate conjunction), det (determiner), noun, prep (preposition), pronoun, quant (quantifier), sconj (subordinate conjunction), scop (scope marker), verb.

b. CS (Constituent Structure): The constituent structure includes a list of surface structures in which the entry can appear.

c. FEATURES: Features that are associated with the entry (one or more).

d. GS (Grammatical Structure): The list of the possible grammatical relations, each followed by an index referring to an element from the CS.

3. FEATURES: Attributes providing information about number, tense, permitted complement types, etc.

4. EX (Example): Example sentence which includes the entry.

Our main concern is the treatment of phrasal verbs in the COMLEX lexicon. COMLEX regards particles as one type of adverbial and uses a separate POS to tag them them (PART). The identification and classification of verb+particle combinations are done using two guidelines. The first is concerned with the distinction between particle and adverbial adjuncts (pure adverbials), and states that in the configuration [V P], where P is either a particle or an adverb, P is a particle only if it does not front. If it does front it should be marked as ADVP,

and also marked with the feature VMOTION[9].

A verb will have the feature VMOTION if it occurs with a locative as a right adjunct. This may be distinguished from verb+particle sequences, such as *He measured up*, by the fact that the adverb permutes (can be fronted). For example:

22a. He walked out.

22b. Out he walked.

23a. He measured up

*23b. *Up he measured.*

VMOTION refers to directional sorts of adverbials which front. For example:

24. He walked in.

25. In he walked.

The idea is that these verbs should take a long list of adverbials, some but not necessarily all of which front, e.g., you might be able to get *Up he jumped* more easily than *Forward he jumped*, but both are covered by VMOTION.

The second guideline covers cases where an adverbial element P is a particle (PART), if the verb takes P alone and also P in a PART-NP complement. For example:

26a. I called up.

26b. I called up John.

26c. I called John up.

However there are exceptions, particularly with meaning changes. For example:

---

[9] ADVPs come in two types: evaluatives and locatives. The latter can front (*He sits here/Here he sits*). For the most part, by locative, we do not mean directional adverbs.

27a. He looked up. (ADVP)

27b. Up he looked. (ADVP)

28a. He looked up John. (PART-NP)

28b. He looked John up. (NP-PART)

*28c. *Up he looked John.*

In the first example *up* in the ADVP sense of *look up* refers to a direction while *look up* in the second example refers to a complex action.

There are cases that are ambiguous between these two structures and can not be resolved using the criteria proposed here. The sentence *He looked up the street* can have both verb+ADVP and V+PART +NP structures depending on the meaning:

29a. I looked up the street and saw John coming down the street.

29b. I looked up the street on the map and could not find it.

In COMLEX each possible combinations of a verb+particle are listed and defined only in the verb entry. Particles are simply listed, and there is no special entry for them which includes a classification and description of their subcategorization frames and their grammatical structure. As a result, the relationship between the verb and the particle is defined in terms of the verb. Under the verb entry the following verb+particle combination (subclasses) are listed in the COMLEX lexicon. For each of these structures the particular particle (ADVL) must be specified and this requirement results in an itemized list of possible combinations for each verb:

1. VERB+PARTICLE: The verbs which are classified with PART take a single particle as

a complement. This particle cannot be analyzed as an adverbial adjunct.

The subcategorization frame of this structure is: NP_PART (the "_" symbolizes the position of the verb).

The verb in a verb+adverb combination can stand by itself, but in some verb+particle combinations, the verb cannot stand by itself without the particle. In the following examples the verb *lined* is ungrammatical without the particle, while *went* in 31b can appear by itself. This indicates that *lined up* is a verb+particle combination, while *went ahead* is not.

30a. They lined up. (V+Prt)

*30b. *They lined.*

31a. They went ahead. (V+ADV)

31b. They went.

Additional evidence that this analysis is correct is the fact that unlike adverbial adjuncts, particles cannot permute:

*32.*Up they lined. (V+Prt)*

33. Ahead they went. (V+ADV)

Phrasal verbs either are ungrammatical when used without the appropriate particle, or, if the verb can occur without the particle, their meaning is different. This can be seen in some verbs by a difference in subject selection:

34a. John carried on.

*34b. *John carried.*

34c. The point carried.

However the sense change does not always occur:

35a. John called up.

35b. John called.

2. PARTICLE+NP

In this structure the verb takes as its complement a particle+noun phrase. The complements of verbs belonging to this subclass could be mistaken for prepositional phrases, but, unlike prepositions of prepositional phrases, the particle and noun phrase may permute, provided that the noun phrase slot is not filled by a pronoun.

36a. He looked the number up. (NP+PART)

36b. He looked up the number. (PART+NP)

36c. He looked it up. (NP+PART)

37a. He looked up the chimney. (VERB+PP)

*37b. *He looked the chimney up.*

37c. He looked up it. (VERB+PP)

The subcategorization frames for this structure are NP_ PART NP and NP_NP PART.

3. PARTICLE+PP (PREPOSITIONAL PHRASE): For verbs of this class there is a strong selectional dependency between the verb+particle on the one hand and the preposition of the prepositional phrase on the other. This dependency is reflected in the fact that omitting any one of the two elements leads to ungrammaticality or a different sense. This construction cannot be analyzed as a combination of a verb+adverb.

38a. She moved in on him.(PART-PP)

*38b. *She moved on him.*

38c. She moved in. (different sense)

39a. He walked around to the station. (ADVERB+PP ADJUNCT)

39b. He walked around.

39c. He walked to the station.

The subcategorization frame for this structure is: NP_PART PP.

4. PARTICLE+NP+PP: In this structure there is a strong selectional dependency between the verb and the particle on the one hand and the preposition introducing the prepositional phrase on the other. This is reflected in the fact that the particle cannot be considered as an optional adjunct as can be seen when it is omitted:

40a. He singled out Mary for the honor (PART+NP+PP)

*40b. *He singled Mary for the honor.*

The particle can appear before or after the NP.

40c. He singled Mary out for the honor.

The subcategorization frames for this structure are: NP_PART NP PP and NP_NP PART PP. It seems this subcategorization frame could be derived from the frames of structure 2 (NP_ PART NP, NP_NP PART) and structure 3 (NP_PART PP) together.

5. PARTICLE-that-S: This subclass of verbs allows the particle to be followed by a clause. The types of clauses permitted are finite complement sentences introduced by *that* and finite wh-clauses. For example:

41a. He pointed out that this was the best approach. (PART+that+S)

*41b. *He pointed that this was the best approach.*

*41c. *He pointed out.*

42a. She figured out what she was doing wrong. (PART+what+S)

42b. She figured out whether they were coming.

42c. She figured out whether to come. (PART-wh-to-inf)

Several samples of COMLEX entries for verbs that subcategorize for particles are given below[10]. For each verb there is a specific list of the grammatical structures in which it can appear and a listing of the specific prepositions, adverbs, and particles it can take. In addition special features of the verb are listed (VMOTION, etc.).

In the first example the verb *cry* can take as complements a prepositional phrase (PP), a particle+prepositional phrase (PART-PP) or a particle (PART) alone. Under each permitted structure the specific lexical items are listed. The possible prepositions are *about, form* and *over.* In the PART-PP structure the permitted particle is *out* and the preposition is *for,* as in *cry out for. Cry* can also take a particle alone as in *cry out.* According to this entry *out* is the only acceptable particle with *cry,* but due to the mechanism used, each grammatical structure is separately listed and thus the particle is listed twice. The same phenomenon is observed in example 44. *Erode* takes only one particle, *away.* Since the verb can appear in both NP_PART NP and NP_PART structures the same particle is listed twice.

A more economical method would be to collapse the structural representation and list the

---

[10]These samples were provided by Adam Meyers (personal communication).

particle only once. *Away* will still be permitted in both structures.

```
43. (VERB :ORTH "cry"
    :SUBC ((PP :PVAL ("ABOUT" "FOR" "OVER"))
        (INTRANS)
        (PART-PP :ADVAL ("OUT")
            :PVAL ("FOR"))
        (PART :ADVAL ("OUT"))
        (THAT-S))
    :FEATURES ((VSAY))

44. (VERB :ORTH "erode"
    :SUBC ((INTRANS)
        (NP)
        (PART-NP :ADVAL ("AWAY"))
        (PART :ADVAL ("AWAY")))
    :FEATURES ((VVERYVING))

45. VERB :ORTH "jump"
    :SUBC ((NP-PP :PVAL ("THROUGH" "OVER" "TO"))
        (PP-PP :PVAL ("ABOUT" "FROM" "ON" "OFF OF" "OFF" "ONTO" "TO"))
```

```
    (PP :PVAL ("AROUND" "ALONG" "ACROSS" "AT" "DOWN" "IN" "FROM"
          "INTO" "THROUGH" "OUT" "OFF OF" "PAST" "OVER" "OUT OF"
          "ONTO" "OFF" "ON" "UNDER" "TOWARDS" "TOWARD" "TO"))
    (NP)
    (INTRANS)
    (PART-PP :ADVAL ("BACK" "AWAY" "OUT")
        :PVAL ("FROM" "TO")))
  :FEATURES ((VMOTION))

46. (VERB :ORTH "jog"
  :PAST "jogged"
  :PASTPART "jogged"
  :PRESPART "jogging"
  :SUBC ((PP :PVAL ("AROUND" "ALONG" "AFTER" "ACROSS" "INTO" "FROM"
          "DOWN" "THROUGH" "PAST" "OVER" "OUT OF" "UP TO"
          "TOWARDS" "TOWARD" "TO"))
      (NP-PP :PVAL ("ALONG" "ACROSS" "AROUND" "INTO" "THROUGH"
          "PAST" "OVER" "OUT OF" "UP TO" "TO" "TOWARDS" "TOWARD"))
    (NP)
    (INTRANS)
    (PART :ADVAL ("BY" "ALONG")))
  :FEATURES ((VMOTION))
```

```
47. (VERB :ORTH "leap"
    :SUBC ((PP-PP :PVAL ("P-DIR"))
        (PP :PVAL ("AFTER" "AT" "P-DIR"))
        (NP)
        (INTRANS))
    :FEATURES ((VMOTION))
```

Let us summarize the approach taken in COMLEX for the representation of phrasal verbs. In general the COMLEX lexicon has broad coverage and is unique in its attempt to handle phrasal verbs instead of just listing them. Whether or not the guidelines used cover the wide range of verb+particle combination, the treatment is consistent and syntactically justified. The approach taken in this project is purely syntactic; there is no reference to semantic features, although it seems that in some cases there is a need to supplement the syntactic structure with additional information to account for different types of phrasal verbs, as will be discussed below.

On the other hand, COMLEX is designed to describe the surface syntactic structure of the entries in the lexicon and therefore does not attempt to deal with the semantic nature of the lexical item. The problem is that syntactic criteria alone does not seem to cover or resolve the ambiguity of phrasal verbs and the present representation do not reflect the semantic ambiguity. This follows in part from the fact that the coverage and concentration of the lexicon is on verbs, nouns, and adjectives, and lexical entries are designed for these

categories alone. There are no separate entries for particles, and they are defined as a constant list of particles which appears only in the entries of verbs.

Basically two criteria are used to distinguish particles from pure adverbials and prepositions. First, particles are distinguished from adverbials in that they cannot be fronted, and second, particles are distinguished from prepositions in that they can permute with the noun phrase. These two criteria, however, cannot cover all types of phrasal verbs, as was shown in Chapter I.

The structural nature of the possible combinations of a verb+particle is defined in two ways: in the entry definition and in the complement structure of the verb. Since the combination is defined in terms of the verb, there is no way to account for the relationship between the verb and the particle from the particle point of view. Some features of the combination can only be defined knowing the particle entry.

In addition, the COMLEX lexicon has separate listings for every possible structure where PART can be found, and this leads to a lot of redundancy. It is possible to formulate the structure so that one subcategorization frame could account for more than one possible structure. The possible subcategorization frames presented in COMLEX for phrasal verbs are:

48. 1. NP_PART

2a. NP_PART NP .

2b. NP_NP PART

3. NP_PART PP

4a. NP_PART NP PP

4b. NP_NP PART PP

5a. NP_PART that-S

5b. NP_PART wh- S

It seems that at least frames 1-4 could be collapsed into one representation that will account for phrasal verb structures with one clause.

49. NP_(NP) PART (NP) (PP).

The collapsed subcategorization frame allows an optional NP on both sides of the particle and allows a PP in the structure with or without an NP structure .

2.2.3 Lexical Semantics

This section presents the problem of semantically representing phrasal verbs in the computational lexicon. First, the nature of the problem is discussed and the importance of an adequate semantic representation is emphasized. Then two approaches to handling semantic disambiguiation are described. The grammar based approach of Ravin (1990), who treats the semantic ambiguity of a specific structure [to V with NP] using grammar-based tools. This work is not concerned with phrasal verbs structure specifically, but nonetheless is relevant to this research, as will be explained below.

Second, Smadja's (1991) approach uses statistical methods to resolve semantic ambiguity. He suggests that statistical approximations such as Mutual Information, t-test, and Collocation-offset can capture semantic relations, and shows how these methods can be successfully used in large corpora to learn about semantically related words without using a grammar.

Since the semantic classification of phrasal verbs is described in detail in Chapter III, thus the main purpose of this section is to discuss the problems that pose concerns specifically for an NLP system, rather than for linguistic theory in general.

2.2.3.1. The Nature of The Problem

If we believe that the goals of computational linguistics are the same as those of theoretical linguistics, namely, to provide useful, testable, and explanatory theories of the nature of language and its relation to human cognition as a whole, then we should attempt to handle lexical ambiguity as an integral part of any language model. To correctly handle constructions that might have ambiguous structures or semantics, we have to rely on an accurate and explicit lexical representation.

Several problems arise in regard to the problem of sense disambiguation. Some disambiguation procedures rely on contextual information. In general, however, contextual factors alone are not sufficient for selecting the correct sense of a word. Often the problem is that the lexical entry does not provide enough reliable pointers to critically discriminate between word senses. Even if all senses are listed, the search process (among all possible

meanings) becomes computationally expensive and often intractable when forced to account for longer phrases made up of individually ambiguous words, as in the case of phrasal verbs. Finally, the chance of creating an exhaustive list of all possible meanings and different usages is slim, new phrasal verbs, are being created in the English language constantly, and it would be impractical to constantly update the lexicon.

The semantic disambiguation of phrasal verbs is an especially difficult task, mainly because of the many-to-many relationship between verbs and particles. One verb may combine with more than one particle, and the very same particle may combine with several verbs. For example, the verb *break* can associated with several particles and convey a different meaning with each of them. The particle *down* can associated with several verbs yielding a different phrasal verb for each combination:

50a. BREAK- up, away, in, down.

50b. break, gun, look, bear, go- DOWN.

This is further complicated by the fact that some combinations of the same verb and particle have multiple meanings, as in *look up*. The literal meaning is coupled with the idiomatic meaning *to search for.* In addition the syntactic structure does not help us disambiguate the different senses or the different types of phrasal verbs in the language.

But the biggest problem of all is determining the semantic compositionality level of the phrasal verb as a whole. The semantic compositionality of phrasal verbs can be described as a continuum. At one end we find literal phrasal verbs whose meaning can be derived from the meaning of its components. At the other end we find combinations which are completely opaque, idiomatic phrasal verbs, whose meaning is idiosyncratic. In the middle of this

continuum we find combinations whose meaning is somewhat predictable from the meaning of its components, but additional knowledge about the nature of the combination is needed to interpret them correctly. These are called systematic phrasal verbs. For some of these combinations the meaning is more systematic and for some it is less systematic. Given the natue of a continuum, it is not surprising that an accurate semantic representation of phrasal verbs in the lexicon is necessary to ensure correct interpretation. Yet constructing this kind of sensitive representation is a complicated task.

Let us consider the important issues, problems, and proposals for creating a reliable semantic representation of phrasal verbs in the computational lexicon. The first problem is exhaustive enumeration of what are regarded as different word senses. There is a need not only to describe all senses of a phrasal verb, but also to try to capture interesting generalizations among the senses of the combination. To do that we must both view the components of the phrasal verbs in isolation and also consider the whole expression in context.

For the purpose of semantically classifying phrasal verbs along a continuum, different levels or perspectives of lexical representation must be specified. A 'static' definition of a phrasal verb will only provide its literal meaning. We need to have some way to describe more 'dynamic' levels of phrasal verbs to account for the interpretation of systematic combinations. One possibility is to use semantic features to describe the general type of phrasal verb the combination belongs to, and in addition provide distinctive features to describe the relationship between the verb and the particle in each case. This proposal, based on Pelli (1976), is explored in detail in Chapter IV. In the next section a rule-based approach

to semantic disambiguation is described.

2.2.3.2. A Grammar Based Approach to Semantic Disambiguation

Ravin (1989) is concerned with building a comprehensive database out of various on-line resources such as a dictionary. In order to do that it is necessary to disambiguate the information found in these sources. The purpose of her research is to build a disambiguation module that analyses a dictionary definition.

Ravin's treatment is important and relevant to our research because her account justifies the usage of "a semantic first notion" for cases when the syntactic structure cannot be determined. According to Ravin, "one way to resolve the syntactic ambiguity is to first resolve the semantic ambiguity that underlies it" (Ravin 1989:260). Exploring this notion is important since in the case of phrasal verbs, syntactic structure cannot fully account for all types.

Ravin's treatment uses semantic features for semantic classification. In Chapter IV I will also propose the use of semantic features to disambiguate different senses of phrasal verbs. Although Ravin is not concerned specifically with the phrasal verb construction in her interpretation of the structure [to V with NP], one semantic type of this structure is the 'phrasal' category which is the phrasal verb usage of an expression such as *invested with authority*. Her treatment of this structure is of special interest.

Ravin discusses the process of identifying and disambiguating the sense of *'with* expressions' of the type [to V with NP] for example:

51a. To fish with a hook.

51b. To strike or hit as if with a bat.

51c. To charge with a murder.

Two types of ambiguities are found. The semantic ambiguity of the elements of the combination and the structural ambiguity of the constituent. In the sentence *The coach hit the player with the bat* the words *hit* and , *bat* have multiple meanings. Disambiguating the sense of the '*with* expression' could help resolve the structural ambiguity (PP attachment) of the combination. Sentences like *The coach hit the player with the bat* can be parsed as syntactically ambiguous between attaching *with a bat* as a modifier of the verb or as a modifier to the noun, with different consequences:

52a. Verb Modifier:

The coach hit the player using the bat. = The coach [hit the player with the bat]

52b. Noun Modifier:

The coach who is holding the bat hit the player. = [The coach with the bat] hit the player.

The '*with* expression' is classified according to its semantic nature as belonging to one of the following categories:

53. a. USE - to fish with a hook.

b. MANNER - to attack with blows of words.

c. ALTERNATION- to fill with air; to mark with bars.

d. PROVISION- to fit with clothes (give).

e. CO-AGENCY/PARTICIPATION- to combine with other parts.

f. PHRASAL - to invest with authority.

The disambiguation module presented in Ravin (1989) proceeds in five steps: First, it identifies the structure in question and the relevant elements (the verb and the head NP) by parsing the string. Then the semantics of each verb and NP is checked in the on-line dictionary and their definitions are also parsed. The semantic relationship between the verb and the NP is determined next using heuristics that contain a set of lexical and syntactic conditions which apply to each category of the definitions collected. For example, the "instrument" heuristics checks for specific conditions that qualify the NP (definition) as an instrument. These heuristics are stated as linguistic rules/ grammar- based conditions. The fourth step determines the meaning of *with* in the combination, and the fifth step assigns the whole expression one of the six semantic categories listed above.

The sixth category is interesting from the point of view of phrasal verb representation. Ravin claims that the PHRASAL category is a purely syntactic function which *with* fulfills in a verb preposition combination. Ravin argues that *with* in such cases simply serves to link the NP to the VERB. In her analysis, Ravin indicates that it is difficult to interpret combinations which are classified as PHRASAL mainly because they might have more than one possible interpretation. She concludes that since the existence of a PHRASAL interpretation is an idiosyncratic property of the verb, there is no general heuristics for solving a conflict of multiple meanings such as in the case of :

54a. to charge with murder. (accuse of)

54b. to charge the battery with power. (refill)

54c. to charge the meal with a credit card (manner)

Another possible analysis could be that in these combinations *with* functions as a

particle, and the reason Ravin is not successful in semantically interpreting it is because she does not consider the verb+particle as one semantic unit. The step in her procedure that considers the relationship between the VERB and the head NP should be modified in the phrasal cases to consider the VERB as a combination of the verb+particle before considering that unit's relationship to the head NP.

The advantage of following this analysis is that PHRASAL will be a semantic category just like the other five categories, and interpretation will be possible where it was impossible before. The only additional information the module will have to include in order to be accurate is some information about the relationship between the verb and the particle in the phrasal combination this will not be possible in case the combination is a figurative one where the meaning is unpredictable in any way.

Ravin (1989) presents an interesting approach to semantic ambiguity resolution. The module she proposes uses grammar-based heuristic which help to determine the relationship between the elements of the [to V with NP] expression and then arrive at a semantic classification. Once a semantic classification is derived, the structural ambiguity can be resolved as well.

Ravin notes in her conclusion that the lack of information about the frequency of word senses provides no principled way to distinguish a primary sense from an infrequent or obscure one. This contrasts starkly with the position taken by Smadja (1991) who uses frequency information to determine semantic relationships between words. His approach is discussed in the next section.

2.2.3.3 A Statistical Approach to Semantic Ambiguity

Smadja (1991), in contrast to Ravin (1989), uses a statistical approach to determine the semantic relationships between words found in large corpora. He relies on the frequency with which two or more words occuring together to determine whether they form a semantic unit. These semantic units are said to be collocations. Several statistical approximations are used to determine the semantic closeness of collocation candidates, among them Mutual Information, t-test, and Collocational-offset.

Mutual Information (MI) is used to evaluate the correlation of co-occurrence of pairs of words. MI was defined in information theory (Shannon 1948) and is used for speech recognition as well as for text analysis. It measures how strongly two events are mutually dependent. *MI(x;y)*, compares the probability of observing word *x* and word *y* together (the joint probability) with the probabilities of observing *x* and *y* independently (chance).When applied to words, MI is defined as follows:

$$55.\ MI(x;y) \equiv \log_2 \frac{p(x,y)}{p(x)\,p(y)}$$

If two words have no dependency, their MI will be null (i.e. $I(x;y) = 0$). The more the words depend on one another the bigger their MI will be. Smadja brings as an example the two words *for* and *example,* which will expectedly have a higher MI than the two words *the* and *up* (Smadja 1991:39). MI does not depend on word frequencies. This is implied by the fact that the denominator for the $p(x,y)$ is the product of the two probabilities. Nevertheless, these probabilities are only estimated probabilities and the estimation is directly dependent on the corpus size and its focus. This can turn into a disadvantage in case the words appear

very seldom or not at all in the corpus or not at all. In these cases, when there is not much evidence for something, it is very hard to know whether it is because it doesn't happen or because you haven't been looking for it in the right way.

The MI statistic is appropriate for assessing similarity between words, but are difficult to use in order to make negative statements about the dependency of two words. The tool which can capture the difference in dependency between words is the t-test (also called t-score).Using the t-test, we are asking which words are significantly more likely to appear after word *x*. If we show that *y* and not *z* is likely to follow *x* then we can form a negative statement about the phrase *(x,z)*. Mutual information is more helpful in identifying associations whereas the t-test focuses on more subtle distinctions.

One possible application of the t-test is in the design of disambiguation rules for a part of speech tagger. In the case of disambiguating the word *to* for example, a t-test can be used to consider words that immediately preceding or following *to*, to determine its correct part of speech and thus its meaning. We discover that words to its left indicate whether *to* is an infinitive marker, while words on its right indicate whether *to* is a preposition. Statistical measurement can provide results that the grammar writer did not foresee writing disambiguation rules. The combination of the tagger with the T-score can help solve this problem.

The third measurement mentioned by Smadja is the Collocation-offset. This measurement provides information on the most frequent structural patterns in which two words occur in a given corpousa. Smadja (1991:181) claims that this analysis allows us to automatically retrieve not only meaningful lexical relations, but also the syntactic relation

of the two words involved. Once these relationships can be determined, the semantic relations are easier to determine too. The relationship is reflected in the relative linear distance between the words. This means that, for example, in given patterns, two words will always be at distance 2 from one another, meaning they will have one word separating them. Also it is possible to have a negative distance which means that one word precedes the other. For example, the two words *back* and *off* were observed 32 times in a(-2) distance, which means that the pattern was *off *word* back* as in the sentence *Take this off my back.* The two words are observed 82 times in (+1) distance, meaning the two words are adjacent, such as in case of *I ask him to back off* (see Table 2.3) The Collocation-offset method is based on the simple observation that some grammatical patterns always hold. For example, when a noun is the semantic object of a verb, then it is principally used after it. In the case of phrasal verbs this observation is a little different. If a particle is used after the verb as in *make up your mind* or *take out insurance,* the particle and not the noun will follow the verb, yet the relationship between the verb and the noun will remain the same. We would like to retrieve this information. In the case of phrasal verbs, the particle can be separated from the verb in the pattern [V NP PART].We know that the verb and the particle can co-occur in a pattern in which the distance between them can be 2 or 3 or 4 etc. In these cases we want to have more information on the most common words that are used in the middle positions, for example, if we knew that *take* and *insurance* appear with the most frequent locator 2, we want to retrieve information about the word that is used in the middle in the recurring pattern *take *word* insurance.* In the case of *take* and *insurance* we would find that the most frequent word used in between is *out.* For cases in which single words are not frequent

enough to be retrieved, part of speech information could be used in order to determine the syntactic relation between the noun and the verb. For example, if the pattern is *w2 *word* w1* and that **word** is an article, then we know that *w1* is used as a direct object of *w2*.

In order to test how effective these three statistical measurements are in establishing the closeness between the components of the phrasal verb, a sample was examined. Two phrasal verbs, *back off,* and *figure out,* were tested for MI, t-test and Collocation-offset. The prediction is that if the pattern is *w1/V, w2*, where *w1* is the verb (e.g. *figure, back*) and *w2*, is the particle (*out, off*) then the pattern *w1 w2* (*figure out* and *back off*) will have high MI and t-test scores. If the pattern is such that *w1* is a noun as in *w1/N, w2*, the MI and t-test will have lower scores. If we run the statistical test without giving any part of speech information the prediction is that the results will be in mid-range (lower than in the case of nd *w1/V* a higher than in the case of *w1/N*).

The statistical approximations were applied to the two pairs of words in a 40 million word corpus and the results show that the predictions were correct as reflected in Table 2.2. A phrasal verb of the pattern *back/V off* has high MI and t-test scores, which support the claim that the two components are semantically related. In the case of *back/N off* the scores are very low and imply that this pattern is rare, and there is no semantic relationship between the word *back* as a noun and *off*. The scores of the pattern *back off* when no part of speech is specified reflect the average scores of the other two combinations and thus that are lower than the maximum and higher than the minimum. The same results are observed for the combination *figure out.* Comparing the results of the two pairs, we observe that *figure out* has higher MI and t-scores than *back off*. This suggests greater closeness between the

components of *figure out* in comparison to those of *back off*.

PATTERNS	AVERAGE - MI	AVERAGE - T-SCORE
back off	1.89	2.79
back/V off	3.89	4.97
back/N off	0.08	0.017
figure out	4.53	6.78
figure/V out	6.39	7.88
figure/N out	1.56	1.56

Table 2.2 MI and t-score Samples

Using the third measurement, Collocation-offset, the prediction is that there should be more cases where the two words of the pair *back off* are separated by the pattern *w1/V, w2*, appear close together than in the pattern *w1/N, w2*. This test will reflect the distance between the verb and the particle . The results for *back off* are shown in Table2.3.

PATTERN	NUMBER OF CASES	W1-W2 DISTANCE
back/N off	36	-2
	1	-1
	1	1
	3	3
	3	4
back/V off	230	1
	1	2
	1	3
	2	5

Table 2.3 Collocation-Offset Samples

There is an overwhelming difference between the distance of *back/N off* and *back/V off*. The words in the latter pattern are usually adjacent. Together with the results of the MI and the t-test, these scores reflect the dependency of the verb and the particle in phrasal verbs.

2.3 CONCLUSION

Macklovitch (1992) examined the grammatical constructions which cause statistical n-gram taggers like PARTS to falter most frequently, and has suggested appropriate solutions to these problems. The model proposed in Chapter IV draws upon Macklovitch's ideas, and his study will also provide the conclusion to this chapter because it highlights both questions and answers to the ambiguity problems.

Macklovitch claims that the main reasons for errors in stochastic taggers are their limited scope, which makes it difficult for them to account for some long distance dependencies, some problems with the tag set used, and some generalizations, both syntactic and semantic, which cannot be captured by a statistical system. Macklovitch would like to automatically detect tagging errors by using the statistical information already at the tagger's disposal and by isolating error-prone contexts through the formulating of linguistic diagnostics in terms of regular expressions over tag sequences. Using Foster's automatic system for tagging English/French (Foster, 1991), Macklovitch lists the most frequent mistakes for that tagger. In his list we find PARTICLES too. It seems that adverbial particles cause 2.9% of all the errors discovered in this corpus. Particles were mistakenly tagged as prepositions 70.8% of the time and as adverbs (RB) 18.5% of the time. Macklovitch claims that taggers like PARTS are not sensitive enough to some semantic relations which determine the appropriate category assignment.

Church (1988) claims 95%-98% accuracy, and while this may be correct for the whole corpus, if we examine the ambiguous words only the accuracy level is no higher than 90%. (see Shaked 1993). When the output of the tagger is the input to a parser this level of

error cannot be tolerated.

Foster's hypothesis is that tagging errors tend to occur in situations where the two best tags on a given token have relatively close scores (competing tags). I found another situation that will result in a tagging error in a statistical tagger, specifically Church's. When the lexical probability of the word points overwhelmingly to a certain tag selection, context is overridden even if it might point to the correct tag choice. Our task is to block or suspend the selection of a tag in certain of these situations. This could be done in the following ways:

1. The tagger can pass ambiguity to the parser in some cases.
2. Potential errors could be reviewed by a human.
3. Foster suggests a statistical algorithm which detects and flags all but 1% of the tagging errors if the difference between the two tag scores is set to 15%. The flagged tokens then will be reviewed by a human.
4. Using linguistic knowledge, formulate descriptions of all these errors-prone contexts in terms of regular expressions over the tag sequences. The linguistic diagnostics will apply to the tagger's output and flag potential errors.

The advantage of using a regular expression is that it become possible to be able to examine a much wider and more complex context than the tagger can while remaining within the same complexity class, i.e., their pattern matching can be done in time which is linear in the length of input. Macklovitch hopes to identify most of the errors as to reduce the burden on the human reviser without requiring all the computational power of the parser.

My proposal, presented in Chapter IV, is similar in some respects to what is

suggested by Macklovitch in the sense that it too involves an automatic method of preventing errors in predictable context. In order to handle problematic cases of Phrasal Verbs we use the statistical knowledge available to examine the context, and we have a general rule which applies after a set of linguistic diagnostics are performed to avoid its overgeneralization . The diagnostics test the context using a regular expression (NP starters) to make sure we are not changing an already correct assignment of PARTS.

Macklovitch experimented to improve past simple/past perfect disambiguation. He formulated linguistic diagnostics which were translated into regular expressions and applied to the tagged test corpus. These diagnostics were intended to identify just the context in which those errors occurred without overgeneralizing to the correctly tagged cases. Macklovitch discovered a trade-off between diagnostics that are efficient in not tagging too many false errors and diagnostics that are effective in not missing many true errors.

Macklovitch concludes that there is a difference in efficiency rates between statistical and symbolic detection methods. We could come up with a mode of cooperation between the statistical and linguistic approaches whereby the great bulk of the work of tagging is done by stochastic methods and the fine tuning (or the error detection) is done with the help of linguistic techniques such as those described above.

Stochastic models provide a strightforward approach to the resolution of linguistic ambiguity by relying on the criterion of distributional frequency when other linguitic criteria have failed to eliminate ambiguity. The use of statistical approximations obtained from large corpora allows the researcher to review previously unavailable information. The fact that words can be associated with frequency probabilities in addition to syntactic and semantic

information enhances the language model and facilitates the processing of large sources of text.

CHAPTER III: THE LINGUISTIC ASPECTS OF PHRASAL VERBS

3.1 INTRODUCTION

This chapter focuses on the theoretical aspects of the phrasal verb construction. The aim is to show the diversity in the structural analysis and the semantical classification of phrasal verbs. This follows from unique properties of the verb+particle combination, which sometimes exhibits a complex verb behavior (one unit) and sometimes phrasal behavior (two separate units). The chapter looks at phonological, morphological, syntactic, and semantic aspects of the phrasal verb construction, in addition to some human acquisition and processing aspects.

No attempt is made to cover all possible treatments but rather to present some different analyses that represent general views in linguistic theory. The main claim of this chapter is that the many theoretical aspects of the phrasal verb construction are not yet resolved and that some of the questions are relevant to the representation and analysis of phrasal verbs in a computational system.

3.2 THE DUAL BEHAVIOR OF PHRASAL VERBS

Four properties have been identified in the extensive literature as being relevant to the question of whether, theoretically, the phrasal verb expression is assigned the status of a word or a phrase. They are: separability, idiosyncratic subcategorization, morphological inflection, and non-compositional semantics.

1. SEPARABILITY: The verb and the particle can occur separately from one another in sentences. This property will argue in favor of considering the components of the phrasal verb two separate elements.

1a. John cleaned up his room.

1b. John cleaned his room up.

In English the position of the verb is fixed and the particle may occupy either the immediate right position or follow the object noun phrase. (In other languages that have a parallel structure, like Dutch and Afrikaans, the position of the particle is fixed, while the position of the verb changes.)

It is obvious that we want to establish some kind of syntactic relation between the two sentences above. Most recent theories (Kayne 1985, den Dikken 1991, Mudler 1992) resolve the structural duality by designating one sentence as exhibiting the deep structure order of constituents, and the other as the surface structure that results from syntactic movement. What kind of movement and what moves and when are debatable and I will present five analyses. What is clear is that we don't want to claim that these two sentences are completely unrelated.

The separability of the verb and the particle is the most salient feature of the phrasal

verb construction and any account should be able to explain the alternative particle positions. The fact that the phrasal verb is separable argues against a one-word, complex verb analysis, unless we advocate processes such as "particle incorporation" and "NP incorporation" (Stowell 1981). These approaches regard the phrasal structure as optionally undergoing some transformations that yield a complex verb form. (See the detailed account below.)

Another feature that ties in with the separability property and argues against the complex verb analysis is the pronominal object. In a sentence like *I looked it up*, the pronoun must come between the verb and the particle, and a sentence like *I looked up it* is ungrammatical. In this case separability is not optional but obligatory, and a syntactic theory will have to explain why. A few modifiers (*right, all, quickly, slowly, etc.*) can also appear before a separated particle.

2a. John looked the information ***right*** *up.*

*2b. *John looked right up the information.*

3a. Yes, he looked it ***all*** *up.*

*3b. * Yes, he looked all up it .*

On the other hand, in some cases phrasal verbs tend to behave like single words in being syntactically more cohesive than ordinary syntactic strings of verb + preposition/ adverb. Fraser (1976) points out that the verb and particle cannot be separated if the verb has undergone action nominalization.

4a. He looked up the information.

4b. His looking up of the information surprised me .

4c. He looked the information up.

*4d. * His looking of the information up surprised me.*

Moreover, the particle cannot be conjoined:

5a. He looked up and over the information.

*5b. *He looked the information up and over.*

Additionally, according to Bolinger (1971), particles cannot be preposed under conditions which normally allow the preposing of directional adverbs. For example:

6a. Up he lifted the weight.

*6b. * Up he looked the information.*

2. IDIOSYNCRATIC SUBCATEGORIZATION: Phrasal verbs can have subcategorization properties that differ from those associated with the verb on its own, and this supports the single unit analysis. For example:

7a. caught = transitive verb

7b. caught on = intransitive verb

As in:

8a. He caught the ball fast.

8b. He caught on really fast.

*8c. *He caught on the problem fast.*

The claim could be made that, once we have new subcategorization frames and new meanings associated with the phrase, we are no longer talking about the same element . *Caught* and *caught on* are two different elements and deserve two different entries in the lexicon. On the other hand we notice that the process of combining verb + particle is a very

productive process. One verb can combine with many particles and these same particles can combine with different verbs yielding new phrasal verbs. In order to capture this productivity we should consider allowing *caught on* to be a derivation of *caught.* The organization, representation, and description of phrasal verbs in the lexicon is one of the goals of this research.

3. MORPHOLOGICAL INFLECTION: Kayne (1985) argues against a complex verb analysis by claiming that should this be the case, then the inflectional affixes would appear on the verbal constituent alone and not on the sequence as a whole as shown in these examples:

9a. The king was countING out his money.

*9b. *The king was count outING his money.*

However the phrasal verb can still serve as a base for word formation rules in the case of a derivational affix:

10. She really is a mixED up kid. (a derived adjective)

11. There were plenty of passERs by. (a derived noun-er suffix)

In English the suffix attaches only to the verb and the particles bear no affixation. Yet semantically the meaning of the derived word involves the meaning of the suffix combined with the meaning of the phrasal verb as a whole. The entire phrasal verb acts as a base for the affixation process. This can be seen, for example, in the derivation of nouns from phrasal verbs. Phrasal verbs undergo zero derivation to form nouns.

*12. He was always a **drop out.***

Likewise in compounding, a nominalized phrasal verb can form the left hand member of a

compound.

*13. We received a **shutdown notice.***

4. NON-COMPOSITIONAL SEMANTICS: Many of the definitions of the phrasal verb rely on the non-compositional semantics of the verb+ particle combination. We find a wide range of semantic classifications for phrasal verbs, ranging from combinations with completely literal meaning to frozen form idioms. Pelli (1976:67) presents four levels of semantic compositionality ranging from compositional to non-compositional. His examples are repeated here to illustrate the semantic range. Each sentence represents a different semantic level.

14. How long before the curtain goes up ? (pure directional: compositional)

15. At this feast he offered up a human sacrifice ! (extended directional: non-compositional)

16. Drink up your beer, why don't you ? (non-directional: non-compositional)

17. Half my actors did not turn up. (idiomatic)

Does the fact that the meaning of a phrasal element does not derive from its parts indicate that the whole phrase should be regarded as one structural unit? Kayne claims that an idiomatic sense of a phrasal verb cannot be an argument in favor of a complex V approach, given the opacity of the notion " idiomatic" and the notion that idioms have the same syntactic structure as non-idioms (Kayne cites Chomsky 1972:169). Any syntactic analysis that relies on semantic evidence to determine structure will have to account for the full range of idiomaticity, from literal to figurative, exhibited by phrasal verbs.

The above characteristics illustrate the ongoing debate about the category that should

be assigned to the structure. Are phrasal verbs a one-word compound type constituent (a lexical category) or a syntactic phrase category where the two words are independent and members of different lexical categories. I will summarize this discussion with Kayne's conclusion: "What we need is a theory that can have 'V PART' mimic V behavior in some cases while prohibiting 'V PART' from mimicking V behavior in others, and the distinction must be a principled one" (Kayne 1985:128).

Below I will discuss the properties of the phrasal verb that make it an interesting and challenging case for linguistic theory.

3.3 PHONOLOGICAL ASPECTS

The issue relevant to our discussion is the prosody of phrasal verbs. In this section I will present an account of the prosody of phrasal verbs from the theoretical point of view. Chapter II already discussed the prosody from an application perspective and showed how prosody in general, and the prosody of phrasal verbs in particular, play an important role in optimizing a speech synthesis system .

From a theoretical linguistic point of view (following mainly Bolinger 1971) we will look at the possible prosodic structures of phrasal verbs and discuss the important factors in determining one stress pattern or another.

Le Roux (1988:48-9) claims that in English, Dutch, and Afrikaans the particle carries primary stress. For example:

18. English:

to drop OUT	*to put OUT*
to break UP	*to take OFF*

19. Dutch:

MEE + maken	*DOOR + halen*
with make	through haul
'to experience'	'to pull through'
	'to scratch out'

20. Afrikkans:

AAN + gee	*OP + klim*
to give	up climb
'to hand (to)'	'to climb up'

Le Roux also points out that in Dutch and Afrikaans this stress pattern is true for compounds where the primary stress typically falls on the left hand, non-head constituent.

Bolinger (1971) provides a much more complex account of the phenomenon. According to him there are three main relevant issues in the prosody of phrasal verbs: The first issue concerns with the way prosody is affected by the existence of more than one word in the construction. English is normally limited to one stress per word, and that limit determines the number of accents that can be assigned to the sentence. The more words there are, the more flexible the speaker can be in changing prosody. The second point is that in the phrasal verb construction there are two words which can be disjoint and can change their position. The order of constituents may affect the meaning, and meaning affects prosody. The third point is that in the case of phrasal verbs we have two or more words sharing semantic features, which Bolinger calls 'a semantic spread out' (1971:45) . Instead of packing a bundle of semantic features into one word, two words will carry the semantic features. A sentence like *He discarded the trash and stowed the bags*, does not sound as good as *He threw OUT the trash and packed IN the bags*, because the semantic contrast in the first sentence is not highlighted, as it is in the second.

As we see, these three factors are tied together and interact to affect the prosody.

Bolinger shows that stress variations depend on the semantic intentions of the speaker and the semantic features of the phrasal verb. He also notes that the fact that some verbs share semantic features with others can affect the stress pattern. For example: the stress pattern of *throw away* changes according to the meaning intended and depending on other verbs in the sentence. Some combinations are impossible, as shown in the example. This example is prosodically ill-formed because it is the particle which conveys the opposite meaning to *keep* and not the verb *throw,* and thus the particle *away* should be stressed rather than the verb *throw.*

*21a.*Shall we KEEP it or THROW it away.*

21b. Shall we KEEP it or throw it AWAY.

22a. Shall we SELL it or throw it AWAY . (throw away = dispose of)

22b. Shall we SELL it or THROW it away. (throw away = discard)

22c. Shall we SELL it or DISCARD it.

A sentence like 21b is possible only when the meaning changes and we want to contrast the two verbs denoting opposite actions.

Phrasal verbs retain a flexible word order and thus a certain freedom of accent. The particle can appear both before and after the complement.

23. He called the man up.

*24. *He the man telephoned.*

It almost seems as if the verb and the particle switch roles, and in these cases the prosody reflects the semantic features embodied in the two phrasal verb components.

25a. Throw that old junk AWAY.

25b. DISCARD that old junk.

26a. They scratched the mistakes OFF.

26b. They ERASED the mistakes.

Bolinger claims that we can use this flexibility as a means of achieving semantic focus. By putting the particle at the end of the sentence and accenting it, we capture the power to make the verb the high point without explicitly degrading anything else. By putting the particle at the end without an accent, we are able to make the verb explicitly redundant. According to Bolinger the following set of examples brings together the major contrasts:

27a. How can they put NIXON over ?

27b. How can they put over NIXON ?

27c. How can they put Nixon OVER ?

27d. How can they put OVER Nixon ?

27e. How can they PUT over Nixon ?

"The first occurs in a setting where acts of political maneuvering are familiar. The second and the third approach the political maneuver from outside: they could be spoken by someone broaching the whole question of politics, the difference being in the relative newsworthiness of the person or the action. The fourth treats Nixon as redundant. The last instance of recertification - the accent displaced onto a syllable that would be meaningless if interpreted as highlighted in its own right implied 'How can such things be?' 'How can they be thinking of such a thing?' " (Bolinger 1971:55).

Another effect which combines both semantic and word order factors is reflected the following examples:

28a. YOU DISCARD that !

28b. YOU THROW THAT AWAY !

There are some limitations in case the speaker wants to increase the number of accents to reflect a level of affirmation or assertiveness. The phrasal verb not only adds an accent of its own but by a rule permits accenting of a demonstrative that is usually not accented in a final position. The "emotional backshift of accent" *The HELL you say!* applies to phrasal verbs in the by moving: stress moves from particle to verb. For example:

29. Turkey ? We have all that we could need plus extra to GIVE away.

30. God, did they CUT them up !

According to Bolinger the choice of intonational pattern depends mainly on the interaction between the structural position and the semantics of the utterance. The difference between particle and preposition behavior is that prepositions are almost never accented in normal speech, whereas particles can be freely accented. The flexible word order in phrasal verb combinations allows stress variation to convey meaning. In most cases, however, the particle is stressed, and thus what looks like identical wording can result in a different structure and meaning depending on the intonational choice of the speaker.

31a. I looked UP the street | and found its location.

31b. I looked the street UP | and found its location.

31c. I looked up the STREET | and saw John.

3.4 ON THE INTERFACE BETWEEN MORPHOLOGY AND SYNTAX

Before proposing an overall structural representation of the phrasal verb construction, we should first explore and evaluate different structural approaches and analyses. There is an ongoing debate regarding the generation of phrasal verbs and their internal and syntactic structure. As we saw in the examples above, the phrasal verb is a verb+particle combination that can be characterized as consisting of a verb and a non-verbal constituent, the particle. The combination exhibits properties of a syntactic phrase on the one hand and of a complex verb on the other. That is, the constituents of the verb+particle combination behave partly as syntactically independent elements and partly as members of a single word-like unit. The dual nature of the construction provides an ideal testing ground for hypotheses concerning the relationship between the morphological and syntactic components of the grammar. The phrasal verb construction has been described as a grey area between these two components.

The aim of this section is to describe the different treatments, emphasizing the need to describe both behaviors (word-like and phrasal) in one account and in one structure. Each view will have to account for dual behaviors of the construction such as: syntactic cohesiveness versus syntactic separability (particle, pronouns, modifiers); morphological operations such as inflectional and derivational affixation; and semantic non-compositionality.

The relationship between the morphological and the syntactic components of a language model have been greatly discussed in the linguistic literature. The phrasal verb construction is described as being on the interface between morphology and syntax. It is

difficult to separate the morphological from the syntactic treatments of phrasal verbs and I do not attempt to do so here. For convenience the two sections are separated under the general title.

3.4.1 Morphological Aspects

Two approaches to the morphology of phrasal verbs are described below: the lexicalist approach and the non-lexicalist approach. An account of the treatment of phrasal verbs in these two frameworks follows the general description of the theoretical issues.

Chomsky's article "Remarks on Nominalization" (1970) concerning the formation of derived nominals in English led to new ways of thinking in the areas of both morphology and syntax. His hypothesis came to be known as the *Lexicalist Hypothesis*. There are a variety of formulations of this hypothesis, but basically the lexicalist approach advocates complete separation between the morphological and syntactic components of the grammar. The three basic hypotheses important to the theory are: *Lexical Integrity Hypothesis, The No Phrase Constraint, and The Lexical Component Hypothesis.*

In its strongest version, Lexical Integrity Hypothesis states that syntactic rules can neither analyze nor change word structure. Each variant of this approach defines "syntactic rules" differently. The No Phrase Constraint says that morphologically complex words cannot be formed by Word Formation Rules (WFRs) on the basis of syntactic phrases. This constraint will prevent syntactic phrases from serving as the basis for WFRs. However, it must be relaxed in some cases, for example to allow compounding. Thus, according to Le Roux (1988:10) "Williams (1981:250) allows exceptional 'headless rules' to form derived

words on the basis of syntactic phrases but calls such headless rules 'sporadic' and 'marked'." Aaronoff (1983:370) permits WFRs to refer to restricted kinds of phrasal information which already exist in the subcategorization frames of lexical entries.

The Lexical Component Hypothesis states that the rules of word structure form a part of a separate component, the lexicon. As Le Roux (1988:12) points out, the basis for this claim is the difference between rules of word structure and rules of phrase structure. Le Roux also points out that in the lexicalist approach the lexicon is taken to be a formally distinct, fully independent "word grammar" which, like sentence grammar, consists of a syntactic component (WFRs), a phonological component, and a semantic component. The lexicon is regarded as being both distinct from and in a feeding relationship to syntax. Taken together, the Lexical Integrity Hypothesis, the No Phrase Constraint, and the Lexical Component Hypothesis constitute a highly restrictive view of the relationship between the morphological and syntactic components of grammar.

In the following section I will present two examples of how the phrasal verb construction is handled within the lexicalist framework. The two analyses I have chosen to describe are Simpson's (1983) lexical V' (V-bar) hypothesis and Selkirk's (1982) dual structure analysis for phrasal verbs in English.

According to Simpson (1983) the English phrasal verb is a V'category which is exceptionally generated by a rule of morphology (WFR). Thus both *look up the number* and *look the number up* are derived from the underlying structure [$[look]_v$ $[up]_p$ $]_{V'}$. On the one hand the construction is assigned the syntactic category of a verb phrase, even though it is

a nonhead constituent, i.e., the particle is not the maximal projection of the lexical category P as required by the X-bar Theory. On the other hand, the deep structure resembles a compound verb structure which is also claimed by Simpson to be generated by a WFR. By assuming that verb+particle combinations are generated by WFRs, Simpson is able account for the word-like properties of the construction (non-compositional meaning, subcategorization, and the ability to serve as a base for other rules of the word formation component).

On the other hand Simpson explains the phrase-like behavior of the English phrasal verb (the construction's syntactic separability and the ability to serve as a base for internal inflectional morphology) by assuming that X' categories formed in the WF component are analogous to syntactic X' categories in that their internal structure is visible. Thus they are accessible to all rules which may subsequently apply to these categories. For example, inflectional suffixes can be attached to the verbal constituent of a verb particle combination. Since the V' category, a verb particle combination, has a visible internal structure, the verbal constituent is available as a base for the application of the rule inserting past tense suffixes. As an independent constituent of the V', the particle can be moved to a position after direct object position to account for the discontinuity of the verb and the particle in a disjoint structure.

There are three problems with Simpson's analysis. Firstly, lexical V' analysis entails generating X' categories by morphological rules. This is, in Le Roux's (1988:65) opinion, not only *ad hoc* but also redundant, since it duplicates the function of phrase structure rules. Also given the assumption that X' categories can be generated by WFRs, it is predicted that other

X' categories apart from V' may be generated in the same fashion, but Simpson does not provide evidence for this.

Secondly, the analysis cannot properly account for the dual behavior of the construction as a lexical item on one hand and as a syntactic category on the other. It does not seem intuitively correct to claim that some V' are generated in the lexicon and others are not without a prediction mechanism that will predict which V' can and which V' cannot.

Thirdly, a device of lexical insertion at higher level nodes is required by a V' analysis, yet is neither justified nor properly described by Simpson. Simpson's lexical insertion device assumes that lexical insertion can occur at the level of non-terminal nodes in phrase structure: "The verb particle combination will enter the syntax with brackets intact. They will be lexically inserted as verb and preposition...under V' " (Simpson 1983:9). In Government Binding (GB) theory, lexical insertion involves X-zero (terminal) levels; thus Simpson's proposal will be an extension of the formal power of the grammar as conceived within the GB framework.

To sum up, Simpson regards the non-compositionality meaning of some phrasal verbs as evidence for the claim that verb+particle combinations are lexical in origin. The WFR assigns a nonlexical category to its output (V'). Simpson distinguishes between phrasal verbs and idioms claiming that verb+particle constructions are not nearly as idiosyncratic as idioms. In addition, idioms are isolated and unsystematic phrasal expressions, while phrasal verbs can be productively and systematically created and new expressions formed can be easily understood.

In my view there are different semantic types of phrasal verbs and some are more like

idioms than others. The division into literal, systematic, and figurative (following Fraser 1976) accounts for all types and explains the regularity of some forms. Non-compositionality cannot be taken as a representative feature for *all* phrasal verbs and thus cannot be taken as a proof for their origin. This is further discussed in section 3.5.

According to Simpson, the verb+particle combinations are generated by lexical rules (WFR in the lexicon). However the combinations are assigned a phrasal category level V'. The V' node, dominating a verb particle combination has to be matched with a syntactically generated V' node whereas the phrase is integrated into a sentential structure. The phrasal verb is assigned a phrasal category to account for its syntactic separability (in accordance with the lexical integrity hypothesis). In addition, the fact that inflectional suffixes are attached to the verbal constituent alone argues in favor of analyzing the combination as two words rather then analyzing them as one phrase.

A second account within the lexicalist approach is that of Selkirk (1982) whose ideas differ from Simpson's in several aspects. Selkirk (1982) accounts for the behavior of the phrasal verb construction both as a lexical and as a phrasal constituent by assigning them a dual structure. She claims that phrasal verbs are assigned both word structure and phrase structure. The formal device proposed by Selkirk to relate the two structural representations is a lexical rule. She advocates two different structures, one for joined and one for disjoint verb+particle combinations. The same sentence, listed again below, will have the following structures:

32a. look up the number.

$[[[look]_v [up]_p]_v' [the\ number]_{NP}]_{VP}$

32b. look the number up.

$[[look]_v [the\ number]_{NP} [[up]_p]_{PP}]_{VP}$

According to Selkirk *look up* in sentence 32a is a compound verb. The particle is assumed to be part of the compound verb. This can account for the syntactic cohesiveness of the verb and the particle, for example the fact that the particle must be deleted along with the verb by the syntactic rule of gapping, which deletes a verb under identity with another verb in the sentence. On the other hand *look* and *up* in 32b are constituents of a syntactic verb phrase where *up* is classified as a preposition since it does not form a compound with the verb.

Selkirk's way to account for the different behavior of the phrasal verb construction and a phrasal constituent is by assigning them dual structural analysis. The notion underlying a distinct representation for the two sentences above is that these two instances are unrelated. It seems to me that a correct structural analysis of the construction would be able to have one account for adjacent and disjoint verbs and particles. Also it is unlikely that *up* in these utterances will be a particle in one and a preposition in the other, especially since the non-compositional meaning and the subcategorization of the two forms are identical. It is not clear what kind of lexical rule will have the power to establish that kind of relationship between what Selkirk considers separate rule systems, namely morphology and syntax. We must look for another way to account for the dual behavior of the expression.

The second approach to morphology is the non-lexicalist approach. The non-lexicalist framework does not have as strict a separation between morphological and syntactic operations as the lexicalist approach. Syntactic rules have the power to analyze the internal

structure of words and major syntactic constituents can appear within word structures generated by word formation rules. For example, within this approach Stowell (1981) argues for postulating a class of WFRs which create syntactic words, that is X-zero categories. Dowty (1979) also distinguishes lexical from syntactic operations within the morphological component and claims that phrasal verbs are created using syntactic operations in the morphology. These two accounts are described below to reflect the non-lexicalist approach to the generation of phrasal verbs.

Dowty (1979) distinguishes between two classes of primitive operations: morphological operations and syntactic operations, presented below in Table 3.1. Both morphological and syntactic operations may be used in either syntactic rules or lexical rules. Dowty believes that there are instances of lexical rules that combine expressions syntactically rather than morphologically so that the derived unit functions as two separate words from the point of view of subsequent syntactic operations. Two such examples are what Dowty calls the verb adjective factitive construction and the verb particle construction, "...kind of lexicalized compound verb though one which typically appears as discontinuous constituents" (Dowty 1979:302).

	Syntactic Rules	Lexical Rules
Syntactic Operations: Traditional syntactic rules.	Phrase structure and transformational rules	Rules forming lexical units of more than one word: verb+ particle combinations and factitives
Morphological Operations: Rules introducing inflectional & derivational morphology	unrestricted, semantically regular, and predictable derivations	Zero derivation and compounds that are partially productive and have less than predictable semantics

Table 3.1: Dowty's Morphological and Syntactic Operations

Dowty does not discuss phrasal verbs at length, yet, like Stowell, this construction exemplifies the interaction between morphology and syntax. He discusses transitive verbal constructions with adverb complements, like *John put the book away*, where the particle can occur before or after the direct object. Dowty differentiates between 'directional verb particle constructions,' which he treats as being formed by compositional syntactic rules, and 'frozen combinations' of a verb and directional adverb where the meaning is clearly non-compositional. Examples 33a and 33b illustrate the former, and examples 34a and 34b the latter.

33a. I put the books down on the table.

33b. He pulled out a gun.

34a. John cleaned the room up.

34b. Will you cut the noise out.

Dowty also distinguishes between the notion of a word and the notion of a basic expression in a Montague grammar. The idiomatic phrasal verb will have to be considered as a single basic expression to get the right result. In other words, non-compositional verb+ particle combinations will be basic expressions consisting of more than one word (as will idioms in general). Dowty points out that the separability feature follows directly from their generation. Once we have specified that verb particle combination rules use syntactic rather than morphological concatenation, it follows that the complex expressions they produce can be discontinuous in full sentences.

Still, Dowty does not explain why 'particle movement' occurs. He suggests Ross' (1967) explanation of 'heaviness' which may be used This states that there is a tendency to order the direct object either before or after the complement of a transitive verb according to the relative 'heaviness' of the two constituents, where the heavier comes last. In this account a particle is lighter than a pronoun: *He looked it up.* However, a particle is as heavy as an ordinary noun phrase and so order is optional

35. He looked (up) the information (up).

In addition, prepositional phrase or adjective complement is heavier than an ordinary NP, but not heavier than a NP+relative clause.

Dowty's model accounts for the duality in behavior of the phrasal verb construction and distinguishes between two groups of phrasal verbs: the compositional ones and the non-compositional ones, though he does not further describe which verbs belong to which group.

The two groups are generated differently to account for their different syntactic and semantic behavior, but all are phrasal verbs. The phrasal verb is generated by using syntactic operations. The compositional ones are derived by syntactic rules, while the non-compositional ones are derived by lexical rules. The lexical rules are used in Dowty's model for less than predictable semantic expressions and partially productive idiosyncratic items. The difference between lexical and syntactic rules is in the domain of application. Lexical rules apply to 'basic expressions' while syntactic rules apply to words. The idiomatic phrasal verbs will be considered 'basic expressions' in Dowty's model, while the compositional phrasal verbs will be considered two words. Yet both forms are generated by syntactic operations.

Another account of phrasal verbs in the non-lexicalist framework is presented by Stowell's incorporation analysis of phrasal verbs, described in his 1981 thesis. Stowell's treatment is similar in many respects to that of Dowty as described above. Stowell distinguishes between X-zero categories that are phonologically interpreted as single words and X-zero categories that are syntactically but not phonologically interpreted as single words. This is parallel to Dowty's notion of a word versus a basic expression.

According to Stowell, the phrasal verb is a complex verb where the particle is actually a complement of the verb appearing as a constituent phrase in V'. The complex verb has an internal XP (complement phrase) for particles.

36. Kevin $[_{V'}[\ _{V}$*[turned] [on]]*$][_{NP}$*the light]]*

This structure will allow Case assignment to an object NP under adjacency since the internal NP will have the status of an incorporated object. The complex verb will be derived by rules

of word formation which belong to a component of extended word formation rules.

Stowell claims that intuitively the verb particle pair functions as a single semantic word, especially in idiomatic pairs. This supports the cases where the verb and the particle are adjacent but leaves open the disjoint cases. We can claim that one structure is transformationally derived from the other, but then we have to assume that syntactic movement rules can apply to subparts of a syntactic word. Stowell assumes the existence of a NP Incorporation rule which is involved in the derivation of the double object construction. He further assumes that in the case of discontinuous phrasal verbs a Particle Incorporation rule has applied to the output of the NP Incorporation. Consider the following examples:

37a. $[_V[_V$ *turned]* $[_{PART}$ *on]* $]$ $]$ $[_{NP}$ *the light]*

37b. $[$ $_V[_V$ $[_V$*turned]* $[_{NP}$ *the light]]* $[_{PART}$ *on]]*

37a is generated by a rule of Particle Incorporation while 37b is derived by the application of a rule of NP Incorporation which applies before the particle is incorporated. The rule of NP and Particle Incorporation can apply separately or simultaneously. These rules are part of 'a component of extended word formation rules', that are different from 'rules of morphology' in their ability to form 'syntactic words' i.e. words that are not phonologically interpreted as single words. Stowell's treatment is very similar to Dowty's, who calls these rules which occur as part of the syntactic operations 'lexical rules'.

Stowell enumerates four advantages of his proposal as follows. First, by assigning the phrasal verb a complex verb category, Stowell obviates the need to postulate a category-specific phrase structure rule such as 'particle movement'. Second, using the rule of Particle Incorporation, Stowell can explain why the adjacency condition on Case assignment is not

violated. After incorporation, the NP is indeed adjacent to its governing verb and Case assigner. Third, because both the continuous and discontinuous phrasal verbs are generated by the same extended word formation rule (Particle Incorporation), there is no need to relate the two structures by means of syntactic movement rule. According to Stowell the discontinuity of the verb and the particle follows from the ordering of the two extended word formation incorporation rules. Fourth, by limiting the generation of the construction to a WFR and not a rule of syntax, Stowell attempts to explain the word -like characteristic of the construction. He claims that phrasal verbs belong to a particular morphological class - the class of native (i.e., Germanic) verbs, as below:

38a. John gave away his money to charity.

*38b. * John donated away his money to charity.*

Le Roux (1988) points out the shortcomings of Stowell's analysis. She claims that the component of extended word formation rules is not well defined. The fact that the Particle Incorporation rule may result in the assignment of different structures to continuous and discontinuous phrasal verbs incorrectly predicts a difference in syntactic behavior between the two types of constructions.

The word-like properties of the phrasal verb could be argued to follow from Stowell's claim that the rules responsible for the formation of this construction are (extended) WFRs. On the other hand, the claim that extended WFRs, such as NP and Particle Incorporation for English, do not create phonological words could serve as a basis for an explanation of the opposite view that claim a phrase-like behavior for phrasal verbs in English.

Let us summarize Stowell's treatment of phrasal verbs as described above. Stowell

assumes that the particle in sentences like *Kevin switched off the light*, where it is adjacent to the verb, is "incorporated" within that verb to form a complex unit. This newly formed verb subcategorizes for an NP and the two together are dominated by V'.

39a. Kevin $[_{VP}\ [_{V'}[_{V}$ *switched off] the light]]*

The main motivation for this analysis is the Case Adjacency Principle (Stowell 1981:113) which requires that, for an NP to be assigned Case, it must be adjacent to the verb. In the example above the NP *the light* is adjacent to the complex V after particle incorporation.

Stowell explains the non-adjacent verb particle structure in the following way: First the WFR of NP Incorporation applies resulting in the creation of the complex verb [switched-the light] then the rule of Particle Incorporation applies to the output and *off* is appended resulting in:

39b. Kevin $[_{VP}\ [_{V'}[_{V}\ [_{V}$ *switched-the light]-off]]]*

The NP has the status of an incorporated object.

As we see, in the non-lexicalist framework the boundary between the syntactic and morphological operations is not as strict as within the lexicalist framework. The dual behavior of phrasal verbs is explained by means of the possible interaction of syntactic and morphological operations.

Other views treat phrasal verbs entirely within the syntactic component. In the next section I will present some proposals for the syntactic structure of phrasal verbs. These accounts treat the construction as a purely syntactic phenomenon.

3.4.2 Syntactic Aspects

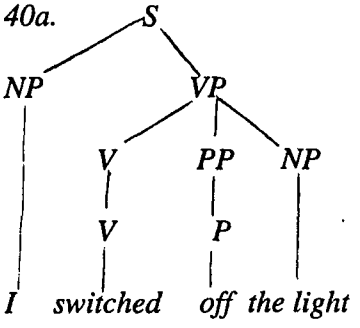
This section presents different syntactic analyses of phrasal verbs. Different approaches phrasal verbs are covered, with an emphasis on discussing as many aspects of the structure as possible (Kayne 1985, Radford 1988, Aarts 1989, den Dikken 1992, Mudler 1992, etc.) The purpose of this section is to illustrate the diversity of treatments and to find a suitable account to incorporate into a computational treatment of phrasal verbs.

The basic positions taken by the studies described in this section are as follows. An initial distinction can be made between analyses based on the premise that the particle forms a lexical unit with the verb that it combines with, and accounts which take the particle to be an independent syntactic head. Another distinction is made between treatments that advocate that in the D-structure the verb and the particle are adjacent (den Dikken 1992), and treatments that claim that in the D-structure the verb and the particle are separated by the object NP (Kayne 1985, Mudler 1992) . Even within the latter view there is disagreement regarding the way the S-structure is derived. Kayne argues for a right NP movement while Mudler claims that it is the particle which moves leftward to incorporate to the verb. The treatments also vary in the way they sometimes consider the particle as a type of preposition. Finally, all but Radford (1988) agree that the Small Clause (SC) is an appropriate structure for phrasal verb although Aarts (1989) advocates this structure only for certain verb+particle combinations. den Dikken (1992) advocates a nested SC structure to account for complex particle structure in one structure.

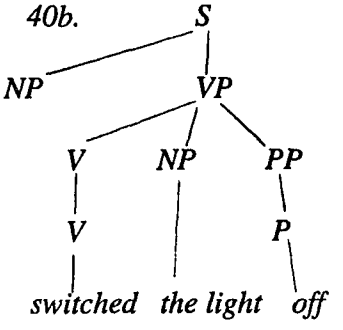
Radford (1988) discusses the phrasal verb structure within transformational theory.

He argues for the following two classes of phrasal verbs.

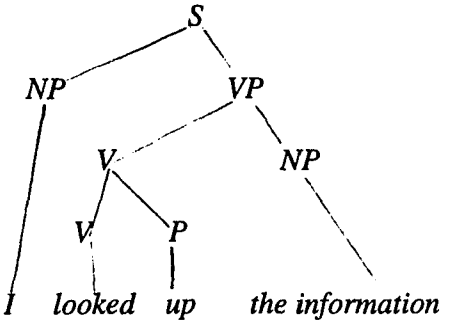
40a.



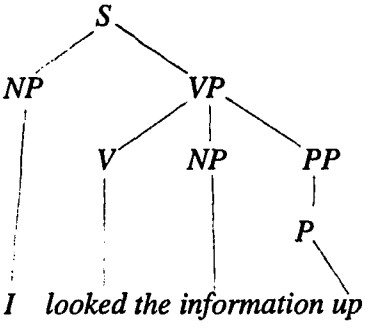
40b.



41a.



41b.



Radford distinguishes between two types of verb+particle combinations. He claims that, while in 40 *off* can be premodified by intensifiers such as *right, completely,* etc, in structure 41 we have a case of a complex verb where the particle can be regarded as word-level adjunction. Radford assigns two different syntactic structures based on syntactic evidence. He does not establish any relationship between the two classes, and Aarts quotes him as saying "to the extent that I envisage any rule relating the two, it's one in which the P originates as part of the PP but is incorporated into the V by REANALYSIS" (Aarts

1989:279). Radford sees the phrasal verb as a combination of a verb + preposition and does not distinguish particles as a separate part of speech.

Kayne (1985) accounts for the construction using his binary branching model[11]. He discusses several possible structures for verb+particle combinations and settles on an SC construction for phrasal verbs, claiming that this structure follows from principles and restrictions in the grammar rather than new rules and structures.

According to Kayne there are three possible structures for 40b and 41b repeated below:

40b. Kevin switched the light off.

41b. John looked the information up.

A: V NP PART

B: [V NP] PART

C: V [NP PART]

Structure A is a flat structure forbidden by the hypothesis that a ternary branching structure is in general unavailable (Kayne 1984). Structure B is not acceptable because the NP is a sister to V' and as such must be thematically autonomous because the NP cannot be assigned θ role by the V'. Non-adverbial NP is not thematically autonomous and thus cannot be a sister to V'. Particles cannot be thematically autonomous because they also cannot be sisters to V'. Thus Kayne advocates C as the correct structure. The verb in this case subcategorizes for an SC which is headed by the particle and whose subject is the NP, as in 42:

---

[11]R. Kayne (1984). Connectedness and Binary Branching. Dordrect: Foris.

42. $[_{VP}\ V\ [_{SC}\ NP\ PART]]$

Kayne calls sentences such as 40 above resultatives. The NP is interpreted as the subject of the particle, and the result is expressed by the SC that is the object of the verb. Kayne and others (den Dikken 1991, Mudler 1992) claim that the SC structure applies to Dutch as well.

43. *switched SUB-the light HEAD-off*

Kayne makes an argument for the SC structure in the following way. He claims that the verb particle construction is comparable to an adjectival SC structure with the same resultative interpretation as in *I switched the light off*:

44. John made Bill unhappy.

In this sentence the head of the SC is the adjective *unhappy*. Kayne claims that by transposition we take the head of the SC in *John looked the information up* to be the particle *up*.

Kayne examines four other properties of the adjectival SC that apply to the phrasal verb construction. One such property is that SC constructions are systematically excluded from having a derived nominal counterpart.

*45a. * John 's considerations of Bill honest.*

*45b. * The looking of the information up took a long time.*

If the [V NP PART] structure is an instance of a SC then both 45a and 45b should be ungrammatical.

Additional evidence that the V NP PART structure is an instance of a small clause involves extraction of the subpart of the postverbal NP. In English extraction of the subpart

of a left branch yields a violation. Therefore, in the case of SC structure extraction will be impossible.

46a. The cold weather has gotten the sister of John quite depressed.

*46b. *Who has the cold weather gotten the sister of [t] quite depressed?*

In the case of phrasal verbs extraction is also ungrammatical, as we see in the following examples:

47a. The cold weather has worn the sister of John out.

*47b. *Who has the cold winter worn the sister of [t] out ?*

The fact that sentences 46b and 47b are ungrammatical proves that the NP is a subpart of the left branch, providing more support for the SC structure.

Kayne explains the ability of the particle to appear after and before the NP by claiming that structure 48a below represents the D-structure order of constituents, while structure 48b is the S-structure order. The underlying structure 48b is achieved by moving the NP to the right, followed by adjunction to V', which results in structure 48c.

48a. V NP PART= *He switched the light off.*

48b. V PART NP= *He switched off the light.*

48c. V[[e_i] PART] NP_i

There is no particle movement and the NP movement is rightward rather than leftward, unlike the treatments proposed by Mudler and den Dikken, which are discussed below.[12]

---

[12] Rightward movement would not be possible according to Kayne's Antisymmetry Theory (Kayne 1993) which allows only leftward movement. According to Kayne (personal communication) he would now tend to support analyses such as den Dikken (1991) which suggest leftward movement of the particle.

Because of the idiomatic character of the verb particle combination in *John looked the information up,* a structure like 48c is less straightforward. Kayne claims that the naturalness of a movement rule is less obvious here than in *John worked off some weight.*

Kayne considers other possible structures, one of which is that the verb and the particle should be taken to be some kind of a complex verb in the base. Here again, the same question comes up as to whether or not we should consider the phrasal verb as a complex unit. Kayne states that although one is tempted to label the verb particle combination as one unit verb it does not seem desirable as we have indications of separability. The following are his arguments against the complex verb analysis:

1. The non-compositional semantics of the combination cannot be an argument in favor of a complex verb approach given the opacity of the notion idiomatic and the notion that idioms have the same syntactic form as non-idioms.

2. If we assume complex verb structure then inflectional morphology should have applied outside the whole complex verb to create the ungrammatical sentence * *John look uped the information.*

3. If a complex verb analysis is assumed then we need to explain the existence of intervening pronouns and some modifiers.

49a. I will look the information up.

49b. I will look it up.

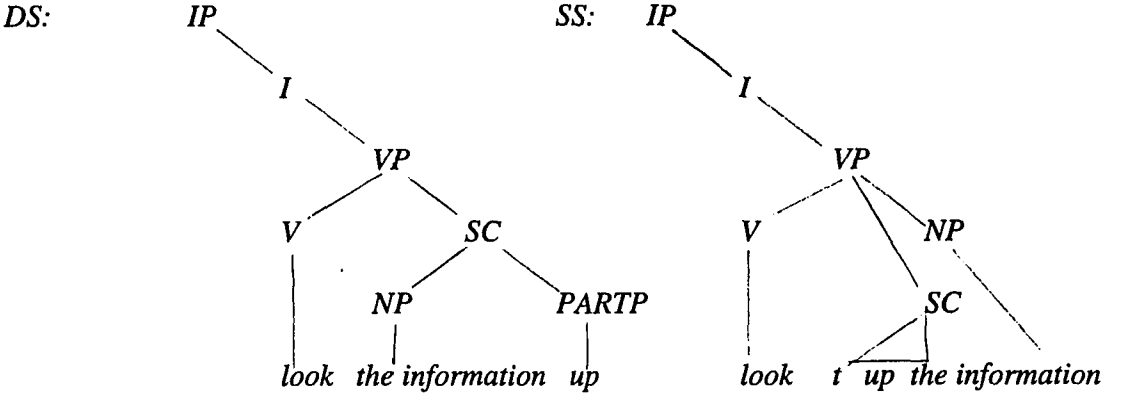
49c. I will look it all up.

*49d. *I look up it.*

Kayne makes no syntactic distinction between sentences such as 40 and 41 above.

He does remark, however, that in the first case the particle expresses a result , whereas in the second the verb+particle has "an idiomatic character" (Kayne 1985:121). In other words, the sentences will be categorized as semantically distinct but will have the same syntactic structure[13]:

50a. *50b:*



As stated above, Kayne rejects the complex verb analysis yet claims that one might be tempted to think about the sequence verb+particle as a one unit verb, although it does not seem desirable to him. Instead there should be a theory which will mimic complex verb behavior in some cases while prohibiting this behavior in other cases in a principled manner.

The GB theory solution proposed by Kayne (1984:128-129) is that the verb+particle sequence can under certain conditions assign θ-role as if it were a complex verb, despite the sequence not being a complex verb. This can be achieved by a percolation mechanism enabling the V' to assign θ-role to its sister constituent in the manner of a complex verb. This

---

[13]Notice again that right adjunction of this sort will not be possible according to Kayne's new Antisymmetry theory which forbids right movement.

θ mechanism does not interfere with the account of the non-complex-verb like behavior of the sequence in other cases. The verb+particle does not belong to the lexical category V but under V' as shown in example 50. Consequently inflectional morphology attaches to the lexical category V alone and not to the combination, which rules out *John look uped the information.*

Following Kayne (1985), Mudler (1992) assumes that the particle in verb+particle combinations is the head of an SC. He also assumes that particles can, in principle, incorporate into the verb at S-structure.

51a. D-structure: *John looked the information up.*

51b. S-structure: *John [looked up] the information.*

One effect of this incorporation is that it excludes right modification because incorporation involves movement of a lexical category. Only the particle can be incorporated with the verb. The modifier *right* in the following example cannot be taken along with the particle, nor can it be stranded.

52a. John looked the information right up.

*52b. *John looked up the information right.*

*52c. *John looked right up the information.*

52d. John looked up the information.

As we see from these examples, the modifier can neither move along nor stay behind by itself. We would expect this structure to be impossible with other modifiers as well, but Mudler finds that the modifier *all* is different:

*53a. *They sent Bill* ***right*** *out the book*

*53b. John sent the people **all** out a schedule*

Mudler claims that the difference lies in the limited distribution of *right,* which can modify non-verbal predicates, while no such restriction holds for *all.* In both Kayne's and Mudler's accounts we find the particle heading an SC. The particle will always be non-adjacent to the verb in the D-structure. But while Kayne (1985) talks about the NP moving and the particle staying in the same position, Mudler argues for a particle movement to join the specifier of V. When this 'particle incorporation' occurs the form becomes one unit and modification is impossible. According to Kayne, the ability of a particle to assign θ-role determines the verb-like behavior of a phrasal verb as a whole, i.e., as one word (Kayne 1985:130). Mudler proposes an actual movement of the particle to the verb, as a result, restrictions apply on what can be inserted between the two words.

den Dikken (1992) is in favor of the latter approach and justifies his choice by arguing that the analysis of particles should be as syntactic heads classified under the category Preposition. According to den Dikken, prepositional phrases can be modified by adverbs such as *right, all* etc, and thus particles pattern as prepositions with respect to modification by these adverbs. Modification will pose a problem for a complex verb analysis as claimed by Kayne (1985) above.

den Dikken defines particles for his purposes in the following way: "In this study the term PARTICLES refers to the class of non-case assigning, argument-taking prepositional elements" (den Dikken 1992:30). He adds, "This is merely a practical delimitation of the object of this study, no necessary theoretical implications being intended" (1992:31).

The essential hypothesis about the features of particles according to den Dikken is

similar in most respects to Kayne's. In particular they agree that: (1) Particles are the heads of small clauses; (2) Particles cannot assign external case; and (3) Particles are non-lexical. den Dikken differs from Kayne (1985) in that he considers particles as prepositions that appear in a nested SC structure. Particle phrases take an additional SC as their complements. The general structure will be:

54. $_{VP}$ *[V [*$_{SC1}$ *SPEC [* $_{XP}$ *PARTP [*$_{SC2}$ *[NP {AP/NP/PP/VP}]]]]*

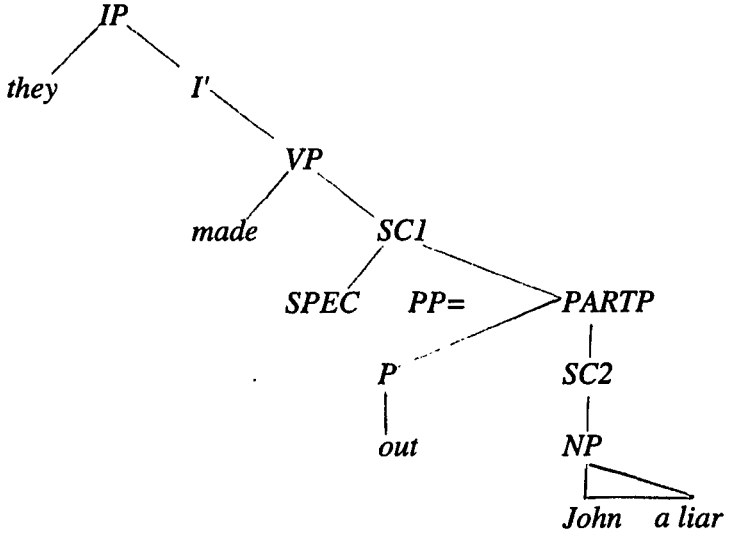
and the following sentences, 55a, and 55b will have the DS and the SS structures shown in 55c 55d, respectively.

55a. They made out John a liar.

55b. They made John out a liar.

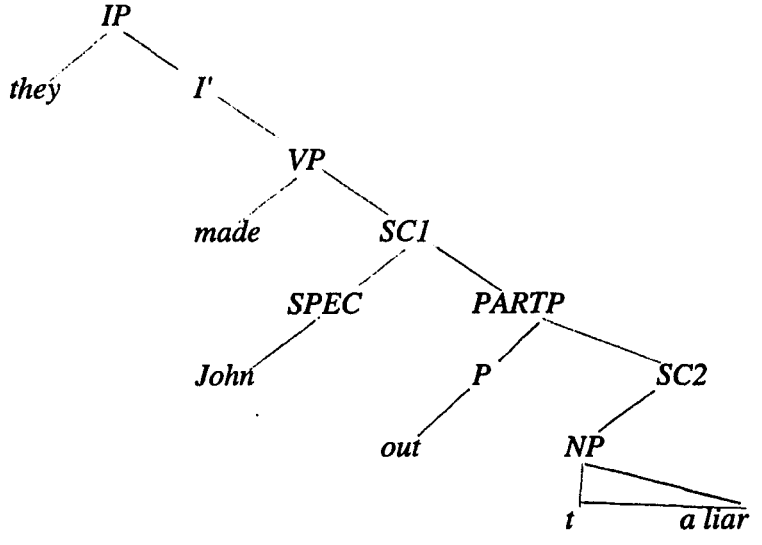
55c.

DS:



55d.

SS:



den Dikken's analysis has several important advantages. Using the nested SC structure results in one account for both regular and complex phrasal verb structures that involve dative structure. This structure can account for verb particle construction in several languages: English, Dutch, German, Afrikaans. Finally, it seems more natural to have an analysis of phrasal verb where the verbs and the particle are adjacent at the deep structure level.

In contrast to Kayne (1985), Aarts (1989) claims that the semantic difference between *switch off* and *look up* results in a syntactic difference as well. He argues that the SC analysis is appropriate only for verb+particle[14] constructions such as *switch off the light,* where there is a genuine subject-predicate relation between the NP and the particle, but is not appropriate for idiomatic constructions such as *look up the information.* Thus he advocates two distinct classes of verb particle constructions which he calls A-verbs and B-verbs.

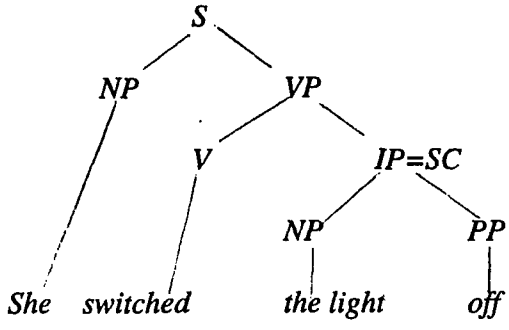
According to Aarts, the difference is that A-verbs such as *switch* take SC complements, while B-verbs such as *look up* (in its idiomatic sense) do not subcategorize for a SC completely but take a regular object and a preposition. It is interesting that this analysis attempts to account for the semantic differences between the different phrasal verbs by having different subcategorizations. In Aarts' treatment, if a verb can appear in both groups (e.g., *look up*) it will subcategorieze for an SC complement in its compositional sense and for a regular object and a preposition in its idiomatic sense.

---

[14]Following Emonds (1972, 1976) Aarts analyzes particles as intransitive prepositions heading a prepositional phrase. Nevertheless, we will refer to them as particles in order to facilitate comparison of Aarts' position with other treatments.

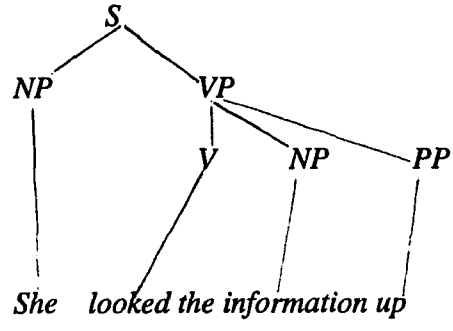
56a.

A-verbs $[_{VP}\ V\ [_{SC}\ NP\ PP]]$



56b.

B-verbs $[_{VP}\ V\ NP\ PP]]$



Aarts' arguments for the A-verb/B-verb distinction are syntactic. He claims that only [NP+PART] complements of A-verbs behave as a constituent and thus should be considered as one. This is assumed by the SC analysis for the A-verbs. As a result [NP +PART] constituents can occur as objects of prepositions:

57. Jim turned the radio off; with the radio off he could finally relax.

This is not possible with B-verbs:

*58. *He brought the kids up by himself; with the kids up he could go on a holiday.*

[NP+ PART] behaves as a constituent when it appears after the comparative prepositions *than* and *as*:

59a. The oven off is less dangerous than the oven on.

59b. The oven off is as dangerous as the oven on.

These last examples show that [NP+ PART] can appear in subject position in sentences with A-verbs but not with B-verbs:

*60. *The kids up is very desirable.*

Given the assumption that units that can be coordinated are constituents, we can test to see if coordination applies to the two types of verbs.

61. Rick switched the lights off and the T.V. off.

Aarts claims that there is a subject predicate relationship between the NP and the particle, which means that the units in question are clauses. B-verbs, however, will behave differently in regard to coordination. Coordination will result in ungrammaticality which suggests that the strings are not constituents.

*62a. *I looked him through and the proposal through.*

*62b. *He sorted the problem out and the cloths out.*

Aarts brings semantic evidence as well as syntactic evidence to bear on distinguishing between two types of phrasal verbs. The semantic test involves levels of idiomaticity, that is, whether the meaning is compositional or idiomatic.

Type A-verbs will have the SC structure and will be of compositional meaning:

63. $[_{VP}\ V\ [\ _{SC}\ NP\ PP]$

Type B-verbs will appear in a flat structure and will have idiomatic meaning:

64. $[_{VP}\ V\ NP\ PP]$

In the A-verb structure, the head V theta marks the SC but not the subject of the SC. The predicate of the clause, in this case PP, assigns the θ-role to the subject NP. The B-verb structure allows the V node to assign θ-role both to the NP and to the PP. Aarts views the PP as a 'quasi argument' - such arguments occur in θ positions. In both instances the case is assigned to the NP by the adjacent verb.

Aarts claims that any sort of complex verb analysis should be rejected since modification is possible. This is true for both types of phrasal verbs, but only when an NP intervenes between the verb and the particle.

65a. I cut the branch right off.

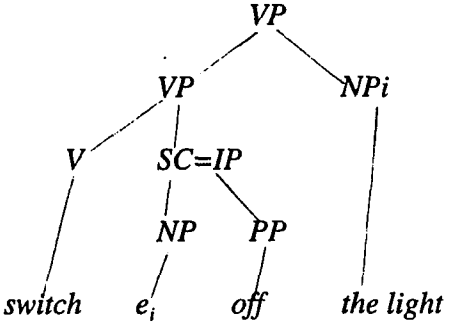
*65b. *I cut right off the branch.*

66a. I looked the information right up.

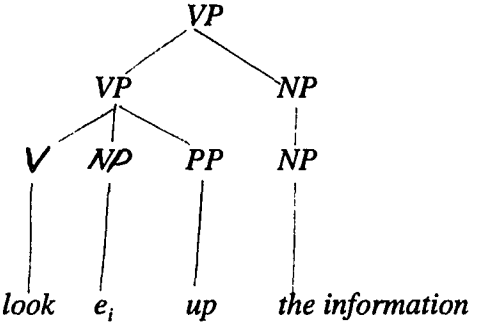
*66b. *I looked right up the information.*

Aarts proposes rightward movement to account for the alternative structure where the particle is disjoint from the verb.

67a. Type-A



67b. Type-B



The NP in each case is adjoined to the VP, as in Kayne (1985), with the difference that in Kayne's analysis the NP is adjoined to V'. Aarts analysis is compatible with Chomsky's claim that adjunction is possible only to maximal projections in non-argument position. In this case adjunction to the SC is excluded in type A because it is an argument position. In both cases the traces are properly governed through antecedent government by the displaced NPs.

Aarts claims that his analysis has an important implication for Case Theory (1989:285). If one assumes that Case assignment takes place in S-structure, then in both above structures (67a and 67b) the head of VP assigns Case to the trace of the moved NP. Because the NPs are moved to a θ assigned position in accordance with the θ criterion, the trace in question has the status of a variable. The movement of the NP in verb+particle constructions is regarded as an instance of heavy-NP-shift, a process that has been argued to leave behind Case marked traces (Stowell 1981:207). Apart from Case, the trace is also assigned θ-role, and both case and θ are transmitted to the NP.

In order to explain why unmarked pronouns cannot appear to the right of a particle in English as in:

*68a. *I looked up it.*

*68b. * I switched off it.*

Aarts proposes that the movement of the NP to the right as an instance of 'move alpha' is always a possible option. The heaviness condition filters out some of the resulting conditions. The heaviness condition says:

69. "A maximal projection XP may appear in an adjoined position after rightward movement across a maximal projection B only if XP is more heavily weighted than B" (Aarts 1989:286).

He presents the following values where XP as any maximal projection that accords with the definition of the heaviness condition:

Heavy XP: 2

Regular XP: 1

Light XP: 0

Kayne (1985) also discusses the issue of heaviness but formulates the heaviness condition in the following way:

70. *"In ...[e]i X NPi where NPi binds [e]i, NPi must be at least as heavily weighted as X weighing: heavy NP=2, ordinary NP=1, pronoun=0, particle=1, right+particle=1" (Kayne 1985: 127).*

While Aarts formulates his condition as a general condition for XP categories, Kayne specifies the weighting of the specific categories involved in the construction. In both cases the particle is heavier than a pronoun, and under this assumption the pronoun cannot appear to the right of the particle.

Another difference between Aarts' and Kayne's notion of heaviness is that Aarts regards the heaviness values not as absolute but as relative within the utterance. For example, if the pronoun is stressed it can occur to the right of the particle and still not violate the heaviness condition as in:

71. Why did you throw out HIM ?

To sum up, Aarts argues for making a distinction between two types of verb+preposition constructions based on syntactic and semantic criteria: A-verbs, which subcategorize for an SC and have a compositional meaning, and B-verbs, which subcategorieze for NP and PP complement and have an idiomatic meaning. For Aarts, the so-called phrasal verbs do not exist and particles are really intransitive prepositions heading a Prepositional Phrase.

3.4.3 Remarks and Conclusion

It is difficult to judge exactly which treatment is the correct one, but certain points in all of these treatments look important and are worth noting. The dual behavior of the verb particle combination is relevant from the syntactic aspect in addition to all the other aspects discussed in this chapter. It seems that even if we claim that an analysis of this combination as a complex verb is inappropriate, we still have to account for its verb-like behavior in certain cases.

Kayne's Small Clause structure seems to explain the syntactic properties of some phrasal verb constructions and their similarity to other resultative constructions, but in some cases the relationship between the verb and the particle seems not to be a resultative one. It is also difficult to determine what kind of movement rule applies here to account for the different positions of the particle and the object noun. The accounts presented differ in their view of which elements move (NP or PART) and in which direction (leftward or rightward), but most agree on the SC overall structure, and in that sense Kayne's contribution is the most important one.

den Dikken's account is based on Kayne's, but takes Kayne's position a step further in suggesting a nested SC structure (more than one SC embedded in an other) to account for regular and dative phrasal verb cases as well. His proposal also abides by Kayne's Antisymmetry theory, which allows only leftward movement. And lastly, den Dikken's proposed deep structure has the verb and the particle adjacent, which intuitively looks more correct.

Aarts interesting contribution is in correlating the syntactic and semantic properties

of what he claims are different types of phrasal verbs. Kayne himself points out semantic differences between phrasal verb types but does not make any correlation between their syntactic and semantic properties. There seems to be a need to establish a relationship between the syntactic and the semantic description of phrasal verbs which will allow wider coverage, have more explanatory power, and be able to distinguish the different types of phrasal verbs and the different verb particle relationships.

The next section surveys the semantic classification of phrasal verbs in the linguistic literature, and tries to point out the important issues, and the questions we should ask when discussing the semantic relationship between the verbs and particles.

3.5 L2 ACQUISITION & PROCESSING

3.5.1 Introduction

The main concern of this work is to show that the analysis of the phrasal verb construction is far from resolved and that in fact in every area of linguistics we find differing views regarding the nature and processing of the structure. Thus, it not surprising that these questions also surface in the field of language acquisition and processing. We do not attempt to solve any questions regarding the acquisition and processing of phrasal verbs, but only to articulate them and point to the main treatments proposed in the literature. The issue of human processing of phrasal verbs is also relevant to this work's interest in machine processing, since it can and should influence the way we design our processing systems. Many questions that are relevant to human processing are relevant to machine processing. For example: the question of representing phrasal verbs in the lexicon has to be addressed because of the need to account for both the human lexicon and the machine lexicon.

3.5.2 Acquisition

The most interesting questions about the acquisition of phrasal verbs concern how this construction is acquired and if there is any correlation between the type of phrasal verb (compositional, non-compositional) and how it is acquired.

In a study conducted by McPartland (1983), Russian-speaking learners of English exhibited what seemed to be a preference for semantically transparent verb-preposition combinations over more opaque phrasal verbs. These speakers used phrasal verbs infrequently in conversation even then incorrectly. Dagut and Laufer (1985) conducted a

preference test for Hebrew/English speakers using Fraser's categories of 'literal', 'systematic', and 'figurative', and found that, in order of frequency, the phrasal verbs used were mainly 'literal', followed by 'systematic', and lastly 'figurative'. They also found a strong preference for one-word verbs over phrasal verbs. A memorizing test produced similar results. McPartland (1989:6) criticizes these studies for not taking into consideration "the complexities of phrasal verbs" and other factors like L1 structure. For example, the fact that Hebrew does not have a structure comparable to phrasal verbs. McPartland's also shows that despite Dagut and Laufer's claim of L1 transfer, there is no evidence that once phrasal verbs are acquired their presence or absence in the L1 affects the access of their meaning.

The studies described above concern which type of phrasal verbs are easier to acquire but do not account for the way in which they are acquired, for instance, whether they are acquired as one phrase or as two units. The main point in all the studies and others mentioned by McPartland (Irujio 1986, Yorio 1989), is that semantic transparency plays an important role in the acquisition of idiomatic expressions in general and phrasal verbs in particular.

3.5.3 Processing Models

Processing models of phrasal verbs follow models for representing idioms in the grammar. The two main approaches are "idioms as a special list" and "idioms as part of the lexicon". In this section, following MacPartland (1989), I will survey briefly the studies done to support each of these access models.

According to the *Idiom List Hypothesis*, idioms of non-compositional nature should

be put in a special list. The implications for processing are that idioms are stored and accessed from a special idiom list which is not part of the regular lexicon. The special idiom list hypothesis can be established only if we have evidence for serial processing . In other words, if evidence for two different sources, a lexicon and an idiom list, is found then we can argue for the existence of a special idiom list. Evidence will be found only if these sources are searched in some serial manner during processing. Since this research interest is in finding the appropriate representation of phrasal verbs, the implication of the special list hypothesis would be that some phrasal verbs, at least those defined as idioms, are represented as one unit entries in a separate list in our grammar.

Clark and Lucy (1975), Grice (1975), Searle and Brannon (1975), and Bobrow and Bell (1973) all advocate serial processing. They claim that processing involves first determining the literal meaning of the utterance, then comparing the literal meaning with the context. If the literal meaning is inappropriate, the conveyed meaning is determined and the utterance is used on the basis of its conveyed meaning. In other words, they all conclude that literal meaning is accessed first followed by figurative meaning.

A variant to this approach is the *Direct Access Model* (Gibbon 1989), which claims that it is not necessary to interpret literal meaning first. Idiomatic meaning can be directly accessed, yet, if the literal meaning use is heard, the figurative meaning is automatically analyzed before it is determined that the literal reading is the appropriate one. Gibbs (1980, 1986) also challenges the "literal first, figurative second" order of serial model.

An alternative approach *The Regular Lexicon Hypothesis*, represents idioms in the lexicon as regular entries and thus for processing purposes are accessed in the same fashion

as all other words. If we find evidence of simultaneous access to both idiomatic and literal interpretations of an idiom then we can claim that idioms are represented in the regular lexicon and thus phrasal verbs will be incorporated into the lexicon in the regular fashion.

The advocates of pure lexical processing (Swinney and Culture 1979) bring evidence that both literal and figurative meanings of idioms are simultaneously initiated when the first word in the string is encountered. Psycholinguistic research shows that ambiguities (such as part of speech ambiguity, semantic ambiguity, etc.) are resolved in context. Some parallel processing might occur until disambiguation ends activation of all possible assignments, whether frequently and infrequently used. Swinney (1982) provides empirical evidence for the autonomy of lexical access hence support for a contextual independent model for lexical processing. Evidence for simultaneous processing does not support the notion of a special idiom list.

McPartland (1989) deals with idioms in general and phrasal verbs specifically. Following Swinney's notions she provides supporting evidence indicating that both literal and figurative meanings of phrasal verbs were accessed simultaneously for both native and non-native speakers of English. This happened even when the context biased one reading over the other. This observation points against the "literal first, figurative second"order of processing.

McPartland (1989) found that non-native speakers and native speakers behaved in a similar fashion. For both groups there was activation of both literal and figurative meanings supporting a simultaneous model of access. And in addition for both groups lexical access was shown to be independent of contextual bias.

McPartland (1989) concludes that figurative meanings of phrasal verbs are lexicalized. The idiomatic meaning of the two words is stored as a unit and therefore, no special idiom list is accessed. However, conclusion still leaves some questions unresolved. It is not clear what is meant by the claim that the two words are stored as a semantic unit in the regular lexicon. McPartland does not specify how the phrasal verbs are listed in the lexicon, for example, whether we have three entries for *gun down* in the lexicon: one for *gun*, a second for *down,* and a third for *gun down* It is not clear what activates the literal and figurative meaning - whether the verb *gun* triggers all possible usages (including the phrasal verb use) or the combination as a whole triggers the phrasal verb meaning. This could be checked by testing the processing of the construction when the verb and the particle are disjoint.

3.6 LEXICAL SEMANTICS ISSUES

3.6.1 Introduction

The semantic analysis of phrasal verbs is anything but straightforward. They are not only syntactically ambiguous, even when identified and assigned a structure one phrasal verb can have more than one meaning. Semantic classification of phrasal verbs is hard because we have many-to-many mapping and even in the cases of one pair we can have multiple meanings, as in the case of *look up*:

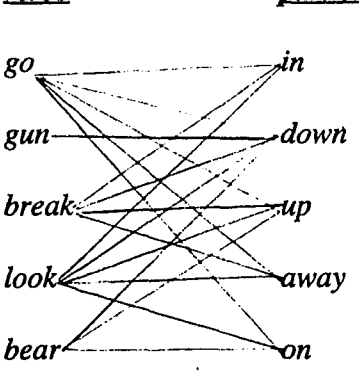
72a. look up = searched *I looked the information up.*

72b. look up = admire *I looked up to you. You were my model.*

72c. look up = looked upwards *Please look up don't look down.*

The list below illustrates the many-to-many relationship described above:

73.



3.6.2 The Lexical Semantics Classification of Phrasal Verbs

It is clear that there are different types of phrasal verbs and that one expression can have more than one meaning. The main concern of this section is to determine how verb+particle constructions can be distinguished semantically from one another, and how the multiple senses of one phrasal verb can be distinguished. The assumption underlying our treatment is that it is crucial for the interpretation of the whole phrase to observe the interrelation of the verb and the particle, especially in cases where the meaning is not compositional. In the following section I will describe in detail the different semantic classifications proposed for phrasal verbs, starting with Fraser (1976), whose basic three way distinction will be adapted, followed by Bolinger (1971) and Pelli (1976).

3.6.2.1 Fraser's Literal, Systematic and Figurative Classes

Fraser (1976) claims that a verb and a particle can semantically combine in more than one way. He characterizes a variety of these combinations in terms of the effect the particles have on the verb. Fraser classifies phrasal verbs into three main groups: literal, systematic, and figurative.

In the literal cases the verb and the particle semantically combine so that the meaning of the whole expression is composed of the meaning of it parts (i.e., the meaning of the verb alone and the meaning of the particle alone). The meaning is thus predictable and can be derived by extracting the lexical information from each of the elements separately.

In the combinations of a systematic nature, the particle appears to cause a consistent change in the meaning of the verb stem. Consider the following examples:

74a. bolt down, drink down, swallow down

74b. store away, cache away

74c. hang up, nail up, screw up

74d. give over, hand over

Fraser claims that, although the notion of systematic cannot be precisely formulated, it is possible to distinguish between cases of consistent process of modification and other cases which result in frozen forms. The latter will be called figurative combinations.

The particle can affect the verb in two ways in a systematic relationship. In the first case the particle can act as an adverbial force, as a resultative. For example, if one *hangs up* a picture, the picture is *up*. This is not true for figurative combinations. For example, if you *look the information up*, the information is not *up*. In the second case the particle can act as a modifier to the verb as in *faded away*.

Fraser suggests that systematic phrasal verbs can also be classified according to their co-occurrence restrictions into those for which the co-occurrence restriction on the verb and the verb-particle combination are identical and those for which it is not. For example, the verb *hand* has the same subcategorization frames as the phrasal verb *hand out,* yet it has different subcategorization frames than *hand over.* Fraser claims that the systematic cases amount to only a small part of the total phrasal verbs in the language and that the unsystematic cases are much more frequent.

Although some figurative phrasal verbs have the same co-occurrence restrictions as the verb alone, there is no way to predict the effect that the addition of the particle has on the interpretation of the verbs. Fraser regards the semantic classification of phrasal verbs as a

continuum of semantic predictability, where on the one hand we have a completely literal, compositional meaning, and on the other hand frozen forms. In between, we have systematic phrasal verbs which can lean towards one classification or the other.

Fraser's general classification of phrasal verbs into three main groups seems intuitively correct. The notion of some systematic influence of the particle on the verb also agrees with the data. Yet it is insufficient in two respects. First, there is a need to see the relationship between the verb and the particle as a reciprocal one, where the verb plays a key role in deriving the meaning of the whole combination. There should be a more explicit account of what the possible semantic interactions are between the two parts of the phrasal verb. The second point is that the three group classification is superficial. It might suffice for the literal and figurative classes, but definitely does not for the systematic class. There is a need to explain the different systematic meanings possible and to describe the semantic patterns of the subgroups of the systematic phrasal verb class. In this way we will be able to account for the predictability of certain combinations and will be able to identify a systematic phrasal verb under certain generalizations derived from this classification. In the next section I will describe several attempts to account for these critical components that are missing in Fraser's analysis.

3.6.2.2 Bolinger's Stereotyping Levels

Bolinger (1971) states the following lexical semantic characteristics for verb particle combinations. "In its core meaning (though not necessarily in the figurative extensions) the

particle must contain two features, one of motion-through-location, the other of terminus or result" (Bolinger 1971:85).

Bolinger claims that the notion of perfectivity can be extended to cover the bulk of phrasal verbs whose meaning deviates from the more or less literal sum of its parts. Still, a more explicit treatment is needed for individual particles. Bolinger says that there is no real borderline between non-aspectual and aspectual uses of the particle but rather a continuum. He offers up as an example the following aspectual levels:

1. The primitive directional meaning, literal or metaphorical.
2. Extended directional meaning.
3. Perfective meaning as manifested in resultant condition.
4. Perfective in the sense of completion or inception.
5. Perfective in the sense of attaining a high intensity.

According to Bolinger, phrasal verbs present a semantic gradient from highly concrete meanings of direction and position to highly abstract meanings akin to aspect. At one extreme *He pulled the boat alongside* tells us nothing but the direction of the motion and the position of the object. At the other extreme *He drew up a list* is purely aspectual. Yet, fundamentally, the two extremes are related.

Bolinger notes that the particle sometimes behaves like the verb ("a verb like quality"). For example the particle *down*, exists as a simple verb and many of its senses are shared with the phrasal verb in which it appears. The verb-like quality accounts for whatever semantic features of movement or activity the phrasal verb incorporates. Bolinger presents the following examples to show an increasing verbness of the particle *down* from extreme

literal to complete idiomatic meaning (Bolinger 1971:95).

75a. Aim the gun down.

75b. Toss me down the top package.

75c. Live down your past.

75d. Spell them down.

75e. Scale down your demands.

75f. Get down the pill.

Bolinger points out that the phrasal verb construction is at the border of syntax and morphology. He raises the question as to what extent the elements of the phrasal verb may be captured in one position or the other to yield combinations that are relatively or absolutely inseparable in meaning, making the combination a lexical item. Bolinger indicates that phrasal verbs are special in that they represent a kind of double layer of compounding. The particles are more or less affixational in nature and the very fact that the combinations pass the definite-noun-phrase-test (see section 1.3.1), classifies them as a compound of some sort.

The first compositional layer is the simple association of a verb and a particle. The second layer is a differentiation within the phrasal verb, related to the varying position of the particle and other factors. Bolinger suggests three levels of stereotyping to account for the different structural and semantic types.

First Level Stereotyping consists of "..loosely connected but nevertheless to some extent ready made combinations that give the lie to notions of a syntax made of an elementary verb and a particle that can join at will" (Bolinger 1971:112). First level stereotyping is a simple combination of a verb proper with a particle where the meaning is

as nearly compositional (Bolinger uses) "additive" as can be. A first level metaphor is one in which the literal meaning of the particle is extended. Literal *up* in *go up* becomes the metaphoric *up* in *load up,* both are first level.

The **Second Level Stereotype** is a phrasal verb that is no longer semantically compositional, and in a second level metaphor the meaning of the whole phrasal verb is figuratively extended. (The difference between the second level metaphor and the second level stereotype is not clear and will be discussed in the evaluation of Bolinger's treatment below.) According to Bolinger second level metaphors make up the bulk of second level stereotypes.

There is also a **Third Level Stereotype** in which the entire verb phrase is frozen. These are idioms, and most have the verb and the particle in adjacent positions. For example:

76a. to strike up the band

76b. to turn over a new leaf

76c. to choose up sides

76d. to take up arms

76e. to keep one's hand in

76f. to put one's foot down

76g. to keep one's shirt on

Table 3.2 introduces Bolinger's three stereotyping levels and provides examples for each category.

	Level 1	Level 2	Level 3
Stereotype	Literal additive meaning: *I came across a field*	The meaning is non-additive: *I came across this finding*	Frozen forms, idioms: *To cry one's eye/heart out*
Metaphor	Literal meaning of the particle is extended : *load up*	The meaning of the whole phrasal verb is figuratively extended.: *Dusk was creeping up on us*	

Table 3.2: Bolinger's Stereotyping Levels

Bolinger brings semantic, syntactic, and phonological evidence to bear on the case of stereotyping levels. He claims that these levels are reflected in the range of closeness between the verb and the particle. Semantic evidence includes the fact that in a dictionary we do not find the sense of *to shape up* in the entry either for *to shape* or for *up.*

Some phrasal verbs allow reduction. This is according to Bolinger the phonological evidence for a verb-particle closeness, for example, vowel reduction in the following cases:

77a. [giddup] *get up*

77b. [g'way] *go away*

77c. [siDawn] **sit down**

Syntactic evidence involves the displacement of the particle. No second level stereotype allows the particle in initial position. This test can determine whether a particle has lost its literal meaning. In other words the phrasal verbs that pass this test belong to the first level.

The following pairs pass the test, allow particle in initial position, and thus have literal meaning and belong to the first level stereotyping.

78a. Down they sat.

78b. Away they flew.

78c. On they came.

The following pairs do not pass the test, do not allow the particle in initial position, and thus have figurative meaning and belong to the second or third level stereotyping[15].

79a. He broke down.

*79b. *Down he broke.*

80a. He gave up.

*80b. *Up he gave.*

81a. They found out.

*81b. *Out they found.*

This test, however, fails to distinguish between first level stereotype and metaphor where the phrasal verb is specialized in some way. Both phrasal verbs belong to the first level stereotype yet when preposed one is ungrammatical; only *set off* keeps its literal

---

[15] In the COMPLEX lexicon (see section 2.2.2.2) this test is used to distinguish adverbs from particles. When the word in question can be fronted to an initial position it is considered an adverb.

directional meaning.

82a. So they set off on one of the longest journeys in history.

82b. So off they set on one of the longest journeys in history.

83a. So they set out on one of the longest journeys in history.

*83b. *So out they set on one of the longest journeys in history.*

Bolinger claims that the second level stereotype resists separation such as displacement by interpolating an adverb between the verb and its particle.

84.	*to catch on (understand)*	**He caught quickly on*
85.	*to take off (depart)*	**The plain took thunderously off*
86.	*to grow up (mature)*	**A child grows quickly up*

Bolinger claims that metaphors will resist insertion of adverbs. In some cases, however, a second level stereotype can move to the third "frozen" level by incorporating idiomatically one adverb where another synonymous one will not be acceptable. For example,

87a. They pressed bravely on.

87b. They carried bravely on.

88a. They carried determinedly on.

*88b. *They pressed determinedly on.*

The illformedness of 88b shows that carry on but not press on has become a third level stereotype.

Let us sum up the particle position argument and its connection to the different

stereotypes. In a second level stereotype that is slightly specialized from the literal meaning, both positions of the particle --before and after the NP-- will be available. A combination will show a greater cohesion between the verb and the particle in its figurative interpretation.

89a. They covered the bodies up. (first level metaphor)

89b. They covered up the crime. (third level metaphor)

Bolinger's semantic classification of phrasal verbs is very broad. He acknowledges the existence of different semantic levels yet does not clearly define them. The notion of stereotyping levels is confusing, since he uses the same terms to define different levels, for example his use of the term 'figurative' to define first level metaphor, second level stereotype and a second level metaphor. There is not enough information to distinguish the first and second levels. Bolinger is talking more about notions than definitions, and the difficulty of defining the semantic continuum of phrasal verbs has already been pointed out. Nevertheless, in designing a lexical representation for phrasal verb in an NLP lexicon we have to represent the different semantic levels in some way as clearly as possible. If we cannot do this to some extent then we will have to create a special phrasal verb list to account for all possible combinations, and this would clearly not be desirable.

In addition Bolinger uses the notions metaphorical, figurative, and idiomatic differently than Fraser (1976), Pelli (1976), and others. What he defines as third level stereotyping can in fact be the class of whole idiomatic expressions, which includes not only the phrasal verb but other elements in the string such as modifiers, for example, *pressed bravely on.* The whole expression and not the phrasal verb alone functions as a one unit

idiom. Fraser, Pelli, and others use the notion 'idiomatic' as synonymous to figurative for phrasal verbs with non-compositional semantics. Bolinger classifies this group as second level stereotype/metaphor. Bolinger uses syntactic tests to establish the semantic stereotyping level, but his tests miss the systematic relationships that are so important for the definition of the relation between the verb and the particle and the predicative power of those generalizations. Bolinger's basic levels may be correct, but for the lexical definition in a computational lexicon we need to be more detailed about these notions and to be able to represent them more clearly.

3.6.2.3 Pelli's Two Level Analysis

Pelli (1976) suggests a two level analysis. The first level examines the relationship between the verb and the particle, based on Fraser's distinction between literal, systematic, and figurative. The second level inspects in more detail, using semantic features, the nature of the particle and how it interacts with the verb to yield the different meanings.

In the first level distinction Pelli distinguishes four groups of phrasal verbs according to the semantic relation between the verb and particle :

The first group includes phrasal verbs of pure directional/motional/location sense. The verb activates the particle to have pure directional-motional or location sense in relation to the meaning of the verb. The particle adds a new directional element to the verb.

90a. How long before the curtain goes up ?

90b. How many steps can you jump up ?

These semantic relations seem to parallel in many respects Fraser's definition of 'literal.'

The second group includes phrasal verbs of modified directional/motional/location sense. Besides the directional/motional/location sense, there is an additional modification resulting from the combination of the verb and particle. This modification is what Fraser calls 'systematic.' The notion that the change to the verb is systematic, partly compositional, or better yet predictable.

91a. As years passed by he forgot her.

91b. You could hear him cry out for help.

While the first sense is determined more by the particle, the sense in this case is determined by the influence of both the verb and the particle and the nature of the relation between them.

The third class includes non-directional phrasal verbs. The particles associated with verbs in this class have lost their directional-motional meaning. The verb in question retains its basic meaning while the particle added to it is mostly contributing a new element to the verb. These are not idioms because the verb retains its basic meaning.

92a. Write out a check for a $100.

92b. They burned down the Temple.

This group corresponds to Fraser's 'systematic' constructions because the overall meaning of the phrasal verb is composed of the meaning of the verb and the meaning of the subgroups, described below.

Finally, idiomatic phrasal verbs consist of idioms, and other expressions of non-compositional meaning in the spirit of Fraser's 'figurative' sense.

93a. I have given in and let her go.

93b. When we broke up I felt alone

The second level distinction is due to the wide range of meaning of verb+particle combinations. One way to account for all possible meanings would be to use semantic features. We can classify the phrasal verbs according to their semantic classes. More specifically, depending on the combination of a certain verb and a certain particle, we can classify what kind of aspectual meaning is conveyed. Thus, one particle may belong to more than one semantic group, but in combination with certain verbs we can predict the aspectual sense of the whole phrase. The second level semantic features describe the way the particles influence the meaning of the verb and the combination as a whole.

Particles function in two main ways: the first may be termed the autosemantic particle, where the particle carries the semantic information needed to interpret the whole phrase, and the second is the auxiliary particle, whose meaning only supplements the meaning of the whole expression. Pelli claims that the particles of the first group have autosemantic function exclusively.

Pelli's subgroups are presented below with an indication of the parallel classification in Fraser (1976). Pelli uses semantic features to describe the subgroups. These features help to understand in what way the particle modifies the verb, the way the verb changes in the presence of the particle, or the way the whole meaning of the combination is changing to become non-compositional as a result of the verb particle relationship.

A. LITERAL COMBINATIONS

Group 1: Directional Meaning - the particle modifies the basic directional meaning of the

verb. The subgroups (1-4) are associated with semantic features (of the particle or the whole expression) and help to clarify in what way the particle modifies the verb. The meaning is compositional.

1. Pure directional-motional particle: *Look around.*

2. Horizontal movement in particles: *Walk up to somebody.*

3. Locational Particles: *Have somebody over.*

4. Directional - motional phrases: *He was called away.*

B. SYSTEMATIC COMBINATIONS

Group 2: Extended Directional Meaning - the basic meaning of the verb is modified by the particle and the whole combination gets a new meaning. The features help to clarify the way in which the particle modification extends the meaning of the verb to create a new meaning for the whole expression which is not entirely predictable.

1. Extension: *This guy is hanging around for nothing..*

2. Intensity: *Years passed by.*

3. Audibility/Visibility: *This secret will pop out some day.*

4. "too much": *Bubble over.*

5. "first": *Go ahead.*

Group 3: Non Directional Meaning - The basic meaning of the verbs (which are non-motion verbs) is supplemented by the particle, which functions as a modifier. The particle does not have its literal directional meaning and the effect on the whole meaning of the expression is again non-compositional. The features help to clarify the way in which the particle modifies the verb and changing the meaning of the combination.

1. Completion:	*Play out your game. Dry out.*
2. Visibility:	*Count out the money.*
3. Repetition, Continuation:	*Talk this matter over.*
4. Stop:	*Shut the lights off .*
5. Begin:	*Turn on the radio.*
6. Degree + :	*Warm up, work up.*
7. Degree - :	*The ice melts down.*
8. Intensity:	*Settle down. Dress her up.*

C. FIGURATIVE COMBINATIONS

Group 4: Idiomatic Meaning - the meaning of the combination cannot be even partially predicted from the meaning of its components. There are no features that will contribute to the interpretation process.

94. I told her off and she left the room.

95. I was all worked up and angry and for a good reason.

96. You always stood up for me.

Pelli's analysis makes a two-way separation which enables us to examine more closely the relationship between the components of phrasal verbs. The first level distinction allows us to generalize regarding the general meaning of the combination and to determine the extent to which the meaning is compositional or not. Only in the case of group 1 is the meaning compositional (literal); in the rest it is non-compositional. Among the other three

groups, the meaning of phrasal verbs in groups 2 and 3 seems to follow some systematic patterns which will make the general meaning somewhat predictable. Group 4 combinations are totally opaque.

The second level analysis allows us to further explore the relationship between the verb and the particle and examine the way in which each component plays a role in deriving the final meaning of the combination as a whole. The use of feature classifications allows us to explicitly describe the interpretation that should be assigned to the verb+particle combinations which belong in that group.

Pelli's classification and use of features is appealing to the needs of a computational lexicon and will provide a basis for the model of representation presented in Chapter IV. We would like to make similar generalizations and show that using Pelli's features can help collapse certain phrasal verbs in a way which will enable us to account for the many-to-many relationships described earlier.

CHAPTER IV: A MODEL FOR REPRESENTING PHRASAL VERBS IN THE COMPUTATIONAL LEXICON

4.1 A THREE-LEVEL MODEL

4.1.1 Introduction

In the following section I will propose a three-level model for representing phrasal verbs in a computational lexicon. The three levels are lexical, structural, and semantic, all of which are integral parts of every lexicon. In the current computational linguistics literature there is no account of a systematic treatment of phrasal verbs in an NLP system. Phrasal verbs are treated either as idioms or listed as a special category, and particles are usually listed as a group (Church 1988, Francis and Kucera 1982). Several studies (Macklovitz 1991, Smadja 1991) do acknowledge problems in identifying and analyzing phrasal verbs in general and particles in particular. However they offer *ad hoc* solutions, and there exists no thorough treatment that covers all aspects of the phenomenon. The contribution of this research lies in (1) presenting a comprehensive picture of the many aspects of phrasal verbs in a computational system, especially their ambiguity, and (2) proposing a model for representing them in the lexicon in a fashion that will facilitate identifying, correctly

analyzing, and interpreting them. The proposal is based on three assumptions which are discussed and motivated below.

The first assumption is that a representation model should reflect the fact that there are several types of phrasal verbs that require different treatments. The justification for distinguishing more than one type of phrasal verb is that there is no single definition that covers all phrasal verbs, as was established in Chapter I. Following Fraser (1976), we will divide phrasal verbs into three semantic classes: literal, systematic, and figurative (see section 3.6.2.1).

The second assumption is that in order to correctly represent and analyze phrasal verbs there is a need to rely on both syntactic and semantic notions. Only by using semantic criteria can we classify phrasal verbs, since there are no coherent syntactic differences according to which they can be classified. Nevertheless there are a variety of structures the phrasal verb can appear in that should be considered.

The last assumption is that to facilitate phrasal verb disambiguation, grammatical as well as frequency information should be considered. In Chapter II I described how statistical approximations can be used to establish relationships between collocations. Information of this sort can supplement linguistic information. MI and T-score, for example, can be used to look for potential verb+particle combinations.

The main questions about the phrasal verb construction parallel those asked by theoretical linguists. The first concern is whether phrasal verbs should be represented in the computational lexicon as one or two units and how the construction should be tagged in each of these cases. We must also ask what grammatical structures phrasal verbs can appear in.

And finally we must be concerned with the semantic classes that phrasal verbs belong to and how to use the representation to minimize semantic ambiguity and facilitate interpretation.

4.1.2 Lexical Representation

When discussing the lexical representation of phrasal verbs several issues come to mind. The first issue to be addressed is the problem of constructing a lexical entry for phrasal verbs, since it is not clear how such an entry should be constructed and what it should include. Tagging processes, which depend on the lexical representation, are our second concern, especially the problem of part of speech disambiguation. Once we identify the phrasal verb the relevant question is what part of speech to assign the words in question and how phrasal verbs should be treated if they are considered idioms.

In Chapter III we determined that there are different types of phrasal verbs and that a verb+particle combination is syntactically and semantically different from the same verb by itself. Because of the non-compositional characteristic of certain phrasal verbs, some linguists regard them as idioms. If we consider some phrasal verbs as idioms then the question is whether we should put all phrasal verbs in a special list or have broader generalizations. There are at least three linguistic positions on the idiom question.

According to the first approach, idioms should be placed on a special list separate from the lexicon and thus are not dealt with in the theory. Weinreich (1969:58) advocates the special list approach. He proposes that each entry in the idiom list will include a string of morphemes, associated phrase marker and a sense description.

The second approach claims that idioms should have regular entries in the lexicon

and should be treated like any other lexical item. Following this approach Fraser 's framework (1976) simply lists all idioms in the regular lexicon. He also suggests marking each idiomatic entry with one of six levels of "transformational frozenness" to account for its semantic flexibility. According to McPartland "...Makkai[16] (1973:11-12). proposes a `computerized lexicon' of `lexemes' and `semantic nests' including frequency counts and dialect markers. Each idiom would be listed as a regular entry with its idiomatic meaning, and cross-references given for each of its constituent words" (McPartland 1989:24).

The third approach says that some idioms should be put on a special list and some should be in the regular lexicon, depending on their nature. Katz and Postal (1963:47) advocate this approach. They differentiate between lexical and phrasal idioms and claim that while the lexical idioms should be handled in the regular lexicon, the phrasal idioms should be listed separately. Phrasal verbs are considered to be lexical idioms.

The question for discussion then is whether we need multiple entries for phrasal verbs and their components or can collapse them into one lexical entry and treat them as idioms. Let us consider three possible options.

The first option is to construct three separate entries:

entry 1. verb = look

entry 2. PP/Adv = up

entry 3. PV [verb+PART] = look up

The phrasal verb is placed in a special entry and is listed separately from the verb and the

---

[16]Makkai, A. (1973). The Cognitive Organization of Idiomaticity: Rhyme or Reason ? Georgetown Working Papers on language and Linguistics. 11:10-29. Washington, DC: Georgetown University Press.

preposition in their regular occurrences. There is no need to list the particle as one of the options for entry 2 because the particle will only occur in the phrasal verb combination and all these cases will be listed under entry 3.

Under this formulation all phrasal verbs are entered in the standard lexicon. This turns out to be a disadvantage since we cannot distinguish the different types of phrasal verbs (systematic, figurative), their subcategorization, selectional restrictions, etc. In addition we do not establish the semantic relationship between the simple verb and its the phrasal verb counterpart because we list them as two separate entries. These relationships are important because they make it possible to interpret the phrasal verb, and predict its meaning where possible, for example, in the systematic cases.

The second possible treatment of phrasal verbs in the lexicon is to construct two lexical entries:

entry 1. V = look

entry 2. PART = up

Following this approach we do not have a special entry for the phrasal verb because everything is captured in regular lexical entries. We indicate in the lexical entry for *up* that it can function as a particle and we use the regular grammar rules and semantic interpretation rules to process the combinations.

There are several problems which cause this formulation to overgeneralize the phrasal verb construction. If we have only two types of entries, we are justifying the claim that the phrasal verb has no status of it own but is simply a combination of a verb+particle which can be derived from the entries of its components. This approach makes it difficult to distinguish

between compositional and non-compositional phrasal verbs since both types are listed under regular entries. In addition, non-compositional combinations will be hard to interpret using semantic rules that apply to additive combinations.

The third option is to construct two entries and an idiom list.

entry 1. V = look

entry 2. PART = up

entry 3. PV [Verb+PART] = look up

In this approach we distinguish the different types of phrasal verbs. We put in the idiom list only the figurative phrasal verbs and treat them as one lexical item with a special non- compositional meaning (entry 3). This combination will have the internal structure, verb+particle, which will be tagged as shown in 1:

1. $PV[_{V}look\ _{PART}up]$

Internal tagging is necessary in order to account for non-adjacent pairs. The literal and systematic phrasal verbs are not listed as separate entries. The possible verb+particle combinations are authorized in the subcategorization frames of the verb at the syntactic level. This formulation will make it possible to account for cases such as *look up* which have more than one meaning. The figurative meaning will be listed in the special idiom list, and the literal sense will be derived from the regular lexical entries of *look* and *up.* This is the option that will be adopted in what follows.

4.1.2.1 Tagging Issues

As described above we have chosen three tags to account for phrasal verbs:

PART = particle

V= verb

PV = phrasal verb

The choice of a special tag for particles separate from prepositions and adverbs, even though lexically most elements appear in both groups, is justified phonologically, syntactically, and semantically, as was established in Chapters I and III. In addition, idiomatic phrasal verbs are classified as one element for the purposes of system processing. The notion of a "word" for computational purposes is different from the normal use of the expression. Shaked (1993) showed how errors can result from following the assumption that words are always separated by spaces in a tagging procedure. Macpherson (1991) also discussed the need to redefine the notion of what a word is for the very same reason, namely, that some multi-word expressions should be treated as one unit.

One problem concerns how to capture discontinuous phrasal verbs. If we assign one tag to the pair in the figurative case, it might cause difficulties when the verb and the particle are not adjacent. Internal labeling, as shown above, solves the problem since the verb and the particle will be labeled V and PART, respectively, whether the two elements are disjoint or adjacent, and for semantic purposes the construction will be considered one lexical unit.

The tag is only one type of information in the lexical entry. Syntactic and semantic information must be included for a complete representation. These topics will be taken up turn below.

4.1.2 Syntactic Representation

The COMLEX Syntax will provide the basis for our proposal for the syntactic representation of phrasal verbs (see section 2.2.2.2). Each verb in the lexicon has several subcategorization frames associated with it that describe the argument structures in which it can appear. Many verbs can associate with particles to yield the phrasal verb construction.

The figurative combinations will appear separately on an idiom list, together with their structural variations. Hindle (1991) suggests a structural template for handling idioms, and this method will be adapted for our purposes as well.

Hindle (1991) proposes disambiguation procedures for the word *to* and notes that hard cases for disambiguation are idiomatic expressions that include this word. He operates under the assumption that in order to know the meaning of an idiom one must first recognize it as a unit. He also claims that despite the semantic idiosyncrasy, the syntactic description of idioms for the most part follows from regular syntactic rules. The tags and structural analysis are the same as for a literal phrase where the meaning is additive (Hindle 1991:9). This suggests that in order to correctly disambiguate the syntactic category of *to*, it is sufficient to recognize it as an idiom since the idiom already includes in its description the disambiguation of *to*. Thus the disambiguation problem in these cases can be reduced to the problem of recognizing idiomatic expressions. Our model facilitates the recognition of idiomatic phrasal verbs because they all appear on a special list. Hindle constructs an idiom list which is added to the parser's lexicon in the format of a template where the whole phrase is described as a labeled bracketing. For example:

2. $_{PP}[\textit{without}\ _{NP}[\textit{regard}\ _{PP}[\textit{to}[NP]]]]$

3. $_{ADJP}$*[next* $_{PP}$*[to last]]*

4. $_{VP}$*[go* $_{PP}$*[to bed]]*

Hindle labels phrasal verbs the same way, for example:

5. $_{VP}$*[look forward* $_{PP}$*[to NP]]*

The recognition of idioms is embedded in the parser so that whenever the parser sees pieces of the idiom, an idiom is recognized.

As part of their special entry in our model, idiomatic phrasal verbs will have a template structure (somewhat similar to a subcategorization frame) that will describe the grammatical patterns that the idiom can appear in. This provides the flexibility that is required for representing disjoint verb+particle combinations. For example, there will be a template for phrasal verbs permitting the NP to appear on either side of the particle. In this template we must indicate that only one NP is allowed. This is done by marking the NP's on either side of the particle with superscript (+) notation. The following rule explains this notation:

6. The template is a general description of all possible patterns. In a given template, it is possible to have more than one element marked as (+). In the actual structure it is forbidden to have more than one element marked (+).

This rule means that the *template [verb (NP$^+$) PART] (NP$^+$)* can potentially have one of the following two interpretations:

7a. [VERB (NP) PART]

7b. [VERB PART] (NP)

To account for literal and systematic combinations we need to consider the syntactical

description of two entries, that of the verb and that of the particle.

Following COMLEX, we propose subcategorization frames in the verb entry listing all the grammatical structures that the verb and the particle can appear in . The compatible particles will also be listed for each verb. Under the particle entry, all possible grammatical structures in which that particle can appear will be listed. Thus the verb entry will include only the relevant patterns for the verb and the particle it can combine with, while the particle entry will contain all possible patterns for that particle. Example 8 shows the COMLEX entry for the verb *cry* with the lines numbered for reference. The first line lists the part of speech (verb) and the spelling of the word. The rest of the lines provide the subcategorization and other information about its grammatical structure . The second line indicates that the verb can take a prepositional phrase (PP) complement with the prepositions *about, for,* and *over.* Line 3 says that *cry* is intransitive, and lines 4 and 5 subcategorize *cry* for the grammatical structure PART+PP, where the particle is *out* and the preposition is *for,* e.g., *The baby cried out for his mother.* Lines 6 and 7 say that *cry out* can be followed by a *that-S* sentential clause.

8.

line 1: (VERB :ORTH "cry"

line 2: SUBC ((PP :PVAL ("ABOUT" "FOR" "OVER"))

line 3: (INTRANS)

line 4: PART-PP :ADVAL ("OUT")

line 5: PVAL ("FOR"))

line 6: (PART :ADVAL ("OUT"))

line 7: THAT-S))

There is a certain redundancy in listing the same particle twice unless there is a semantic difference between the combination of the verb and the particle in the verb+particle+PP and the combination verb+particle+that-clause. This redundancy can be eliminated by listing *out* every time a new grammatical structure is mentioned only when the particle has a different sense. We can collapse the grammatical representation while still including all the possible subcategorizations. In the case of the verb *cry,* the structural description can be reduced to the following general template:

9. VERB PART (PP^+)($THAT\text{-}S^+$)

PART=out

PP=for

This formulation stats that the verb has an option of taking a particle+prepositional phrase or a particle+sentential clause. We employ here the same notation and the same rule used in 6 to imply that only one element marked (+) is allowed in a given sentence. This notation is understood as "or" by the computational system.

My proposal for representing the syntax of phrasal verbs is based on the COMLEX notation, but it advocates generalizing rather then detailing all possible grammatical structures. I propose to list more than once in one entry only those particles which have different senses. Another feature that is missing in the COMLEX analysis of phrasal verbs is the subcategorization for an SC. In section 3.4.2 we presented evidence that the SC structure is preferred for phrasal verbs. Although there is disagreement concerning the

internal structure, the derivation, and the permutation of particle and noun phrase, it is generally agreed that the SC structure should be assigned to this construction (Kayne 1985, den Dikken 1991, Mudler 1992, and others). Thus, as one of the possible grammatical structures for particle construction should be PART + SC (see Chapter 2 section 2.2.2.2). This specification will allow the particle to combine with a SC and will account for sentences such as:

10a. They made out John liar.

According to den Dikken (1991) and Mudler (1992) the structure assigned to 10a is:

10b. $[_S NP\ [_{VP} V\ [_{SC1}\ Spec\ [_{XP}\ PART\ [_{SC2} NP\ [NP]]]]]]$

The first NP in the second SC can move to Spec position in the first SC yielding:

11a. They made John out a liar.

The structure assigned to 11a will be:

11b. $[_S NP\ [_{VP} V\ [_{SC1}\ NP_2\ [_{XP}\ PART\ [_{SC2}\ e_i\ [NP]]]]]]$

4.1.3 Semantic Representation: Semantic Groups and the Relationship Between the Verb and the Particle

The semantic analysis used for the representation model is based on Fraser (1965, 1976) and Pelli (1971). Fraser (1965) distinguishes between a systematic and a non-systematic relationship between the verb and the particle, determined by whether the particle can be omitted without any significant change in meaning. The literal relation (as in: *pour out, hang up,* and *hand over*) and the completive relation (as in *tire out* and *mix up*) are

classified as systematic. In figurative relations (e.g., *look up (information), knock out, show off*) the meaning of the whole combination cannot be derived from the meanings of its components.

Pelli describes four groups of verb+particle combinations. In group 1, the verb activates the particle to have a pure directional-motional-locational sense combined with the meaning of the verb. The particle adds a decisive new directional element to the verb it is associated with. The particle is autosemantic, that is it contains the directional-motional-locational meaning of the whole expression which is additive from the individual meanings of the verb and the particle. For example, in the following sentences the particle *up* denotes directional motion towards higher level.

12a. The curtain goes up.

12b. Can you jump up the steps ?

This meaning is activated because *go* and *jump* denote action only, without any directional meaning. The particle adds an entirely new element to the verb.

In group 2, the meaning of the expression is extended motional-directional meaning. The particle functions as an auxiliary, modifying the verb and supporting it in some way. In the following sentences the verbs *offer* and *raise* are more restricted in their basic meaning. The sense of directionality already exits in the verb in some way but the particle intensifies the basic meaning of the verb.

13a. They offered up a human sacrifice.

13b. They raise up Maria.

The way in which the particle modifies the verb is reflected in the features associated with

the particle. Different particles can have different features associated with them.

Pelli (1976:76) uses five features to denote the possible senses of the combination in group 2 (see Chapter III section 3.6.2.3). For our model we will be using three main semantic features for verb+particle combinations of group 2:

1. Extension - the particle extends the meaning of the verb, usually in the directional-motional sense.

2. Intensity - the particle intensifies the meaning of the verb by virtue of reduplicating the verb's meaning.

3. Audibility/Visibility - the particle combines with the verb to accentuate 'audibility' or 'visibility' of the object.

It seems that the two features "too much" and "first" are rarely needed to account for phrasal verb meaning. The senses conveyed by these features can be included under the feature of "intensity."

In group 3 verbs retain their basic meaning and the particle lose their directional-motional meaning. The meaning of the particle is figurative and cannot be predicted. The meaning of the whole is additive but one needs to know the features associated with the particle to determine it. For example in the sentences below the particle *up* cannot be activated in its directional sense due to the nature of the verbs *drink* and *bust.*

14a. Drink up your beer.

14b. Don't try to bust up my marriage.

As we said, the relationship between the verb and the particle can be captured using

semantic features.

According to Pelli there are eight possible features for group 3 combinations. Since it seems that there is some repetition in the senses, and in order to simplify the set of features used, the features are collapsed in our model yielding the following five main features:

1. Completion - the particle conveys a completive meaning 'to the end', 'fully' such as in: *it burned all up* and *he was hunted down.* This feature will cover the feature "stop" mentioned separately by Pelli.

2. Visibility - the particle acquires the meaning of 'bring to light', 'make appea'r, 'make clear', such as in: *count out the money* and *figure out the answer.*

3. Repetition, Continuation - the feature denotes two possible relations. a. by virtue of the particle the action of the verb is repeated again and again, and b. the action of the verb acquires a continuous nature such as in: *talk the matter over* and *he hums along.*

5. Intensity - the particle intensifies the meaning of the verb such as in: *you must settle down* and *slow down please.*This feature will include the feature "degree+" "degree-", which were listed separately according Pelli.

In group 4 the verb and the particle combinations cannot be taken apart semantically. The meaning is figurative/idiomatic. For example:

15a. All my models did not turn up today.

15b. We have been brought up so far apart.

Each particle can have several semantic features associated with it. This explains the fact that a given particle can combine with several verbs, yielding different senses for each

one. The use of features enables us to account for the fact that one word (in this case the particle) can have multiple meanings depending on its environment (in this case the verb it combines with).

In the semantic representation of phrasal verbs in a computational lexicon, one must account for the multiple senses and represent them in a way that enables the system to identify them. In the particle entry we list all the possible meanings that a particle can have. For each sense we list the group number and semantic features associated with it. This is important since there are cases where the same feature can appear in more than one group. For example the feature 'intensity' can appear with group 2 and group 3 combinations. The group number is a source of information about the relationship between the verb and the particle, and it helps interpret the features associated with the combination. In the verb entry, only the particles that can associate with the verb are listed. These particles are labeled with the semantic features that are relevant to that particular verb+particle combination. It is possible that one verb will have the same particle associated with it to convey different meanings. In that case the particle must be listed more than once, and each time it must be accompanied by the relevant feature or features.

Let us consider the following sample lexical entry for the particle *up*:

16. UP: PART

group1

group2 [audibility/visibility] *He showed up at the last minute.*

group3 [completion] *Please sew up this pocket.*

group3 [visibility] *She hunted up a screwdriver to fix the handle.*

group3 [intensity] *The coffee needs to be warmed up.*

For group 1 no features are listed yet the group number can appear alone in both the particle and the verb entries to indicate that the literal sense (directional-motional-locational which is additive) is possible as well. Most particles can have a literal meaning, but according to Pelli the particle *ahead* cannot have a 'group 1 sense.' Thus the combination of any verb with the *ahead* particle will not have that sense.

A sample of the semantic representation of the particle in a verb entry appears below for the verb *cry*. The interpretation of this representation is that the verb *cry* can appear with the particle *out*. The semantic relationship between the particle and the verb is associated with group 2. This means that the basic meaning of the particle is directional, but it also modifies the verb in some way. The fashion in which the verb is modified is conveyed by the feature [audibility/visibility], which is associated with the particle in this combination.

17.

cry [PART]

PART=out [group2- audibility/visibility]

4.1.4 Summary

We have presented a three-level model for representing phrasal verbs in a computational lexicon. The model and the treatment suggested in this research are summarized in the Table 4.1 below:

LEVEL / TYPE	LITERAL	SYSTEMATIC	FIGURATIVE
LEXICAL	List the verb and the particle in separate regular entries in the lexicon. **cry = verb** **out = PART**		Special list, one tag to the whole phrase + internal tagging **PV[$_V$look $_{PART}$ up]**
STRUCTURAL	List all possible patterns in each particle. List only the relevant pattern in the verb entry. List the particle only once for each sense. **cry PART (PP$^+$) (that-S$^+$)** **PART = out PP= for**		List all structural patterns in a template: **look (NP$^+$) up (NP$^+$)**
SEMANTIC	In each particle entry, list all features and group associations that the particle can appear in. In the verb entry list only the ones that the verb can combine with as in: **cry PART** **PART= out** **[group2 audibility]**	The meaning of literal expressions is additive. They are associated always with [group1] in the directional-motional-locational sense.	List the idiomatic meaning of the combination as a whole: **look up = search**

Table 4.1 The Representation Model

4.1.5 The Use of Statistical Information

As we see, the representation of phrasal verbs is complicated and far from straightforward. Frequency information can be used to enhance the representation in several ways. First, it can help determine lexical, structural, and semantic co-occurrence. If it is observed that there is a tendency for certain tags or certain grammatical structures to occur in high frequency with each other, it means that the probability of finding certain words together is high, and we can predict which words are likely to be phrasal verb combinations. In addition, frequency information can help us determine the semantic co-occurrence between a verb and a particle of a certain semantic type. We can measure the probability of a verb combining with a particle of a specific semantic group and the probability of its being associated with certain features. This can help us to predict the sense of the combination in addition to predicting senses of new combinations in the language. Frequency information can also help determine the most frequent pairs in a restricted corpus. There is no need to have a large lexicon in a restricted domain. We can build entries only for the cases that exist in the corpus. Finally frequency information can be used for disambiguation purposes. If we use frequency information to disambiguate between possible assignments we rely on what is probable and we might make a mistake. Yet, we use this information to assist us in making a decision when there is not enough grammatical information about the combination. Relying on frequency information proved to be accurate most of the time, especially in a restricted domain (Church 1988, and others).

In section 2.2.3.3 we discussed the statistical tools that are used in computational systems for collecting frequency information, e.g., mutual information, t-test, and

collocation-offset. In addition, we have described statistical taggers (Church 1988, De Rose 1988) which successfully use statistical approximations such as lexical and conditional probabilities to derive the correct part of speech assignment.

In the next section we concentrate mainly on problems of tagging phrasal verbs. The advantages and disadvantages of using a statistical system are discussed. Also some improvements are suggested and tested to justify the combination of linguistic and statistical information to facilitate disambiguation.

4.2 A PROPOSAL FOR PROCESSING PHRASAL VERBS IN A STATISTICAL TAGGER

4.2.1 Introduction

In section 4.1, we presented a model for lexically, structurally, and semantically representing phrasal verbs in a computational lexicon. We also advocate the use of frequency information to improve language representation and disambiguation. Natural language processing (NLP) and AI in general have focused on building rule-based systems with carefully handcrafted rules and domains of knowledge. But in recent years there has been a growth in the use of probabilistic methods for NLP in general and for tagging in particular. This research supports the claim that both linguistic and statistical approaches to NLP will benefit from a combination of the two fields and that in fact these approaches share some common notions.

It is important to point out that 'frequency' and 'distribution' are terms which are used interchangeably in both statistical and linguistic distributional approaches towards natural language. Harris' (1951, 1970) distribution theory is based on these notions. He uses frequency counts to discover the feature set and word classes of a language (all attributes which a language makes reference to in its syntax) and claims that in fact these features license the distributional behavior of lexical items. If a word/word class poses some features or these features license a specific distributional behavior, it means that we can predict its environment, behavior, etc.

In this section we will concentrate on the task of tagging phrasal verbs and make the claim that a statistical approach to tagging (such as the tagger PARTS (Church 1988)) cannot

sufficiently handle phrasal verbs. Shaked (1993) tested alternative treatments of the problem and improved results by supplementing the tagger with linguistic information and by using the representation suggested in the lexical level of the model (section 4.1). The experiment reinforced the notion that there are different types of phrasal verbs and that they should be represented and processed differently.

4.2.2 The Experiment

The experiment described below examined the current performance of the stochastic tagger PARTS (Church 1988) in handling phrasal verbs. The results of the experiment show that the difficulty in handling phrasal verbs in the tagger arises from the statistical model used. A solution to improve the tagger's performance involves a change in the definition of what counts as a word for the purpose of tagging phrasal verbs.

The basic assumption underlying stochastic tagging is the notion of independence. Words are defined as units separated by spaces and then undergo statistical approximations. As a result the elements of a phrasal verb are treated as two individual words, each with its own lexical probability (i.e., the probability of observing part of speech given word) An interesting pattern emerges when we examine the errors involving phrasal verbs. A phrasal verb such as *sum up* will be tagged by PARTS as noun + preposition instead of verb + particle. This error influences the tagging of other words in the sentence as well. One typical error is found in infinitive constructions, where a phrase like *to gun down* is tagged as INTO NOUN IN (a prepositional *to* followed by a noun followed by another preposition).Words like *gun, back,* and *sum*, in isolation, have a very high probability of being nouns as opposed

to verbs, which results in this misclassification. However, when these words are followed by a particle, they are usually verbs, and in the infinitive construction, always verbs.

The hypothesis is that the error appears to follow from the operation of the stochastic process itself. In a trigram model the probability of each word is calculated by taking into consideration two elements: the lexical probability (probability of the word bearing a certain tag) and the contextual probability (probability of a word bearing a certain tag given two previous parts of speech). As a result, if an element has a very high lexical probability of being a noun (for example: *gun* is a noun in 99 out of 102 occurrences in the Brown Corpus), it will not only influence but will actually override the contextual probability, which might suggest a different assignment. In the case of *to gun down* the ambiguity of *to* is enhanced by the ambiguity of *gun*, and a mistake in tagging *gun* will automatically lead to an incorrect tagging of *to* as a preposition. It follows that the tagger will perform poorly on phrasal verbs in those cases where the ambiguous element occurs much more frequently as a noun (or any other element that is not a verb).

The tagger will experience fewer problems handling this construction when the ambiguous element is a verb in the vast majority of instances. If this is true, the model should be changed to take into consideration the dependency between the verb and the particle in order to optimize the performance of the tagger.

4.2.2.1 The Data

The first step in testing this hypothesis was to evaluate the current performance of PARTS in handling the phrasal verb construction. To do this a set of 94 pairs of

Verb+Particle/Preposition was chosen to represent a range of dominant frequencies from overwhelmingly noun to overwhelmingly verb for the first element of the pair and variety of particles and prepositions to represent the second element in the pair (see Table 4.2). For each pair twenty sample sentences were randomly selected using an on-line corpus called MODERN, which is a collection of several corpora (Brown, WSJ, AP88-92, HANSE, HAROW, WAVER, DOE, NSF, TREEBANK, and DISS) totaling more than 400 million words. The two elements of the pairs were adjacent in all sentences examined.

The focus of this experiment was mainly on the verbal head of the phrasal verb, and thus the evaluation was done on the basis of the correct or incorrect tagging of that element, e.g., *gun* in *gun down*. In the current version of PARTS there is no systematic way to differente between prepositions and particles, and phrasal verbs are not treated as a special category, thus it was impossible to evaluate the performance for these tagging tasks.

These sample sentences were first tagged manually by a human tagger to provide a baseline for comparison, and the manual tags were assumed to be 100% correct. The sentences were then tagged automatically using PARTS in order to evaluate PARTS performance for the very same task. A third approach to the tagging was also explored in order to test if a simple solution could solve the problem. This was done by tagging the first element of all pairs as a verb. The accuracy of the three tagging approaches was evaluated.

4.2.2.2 The Results

The results of the experiment are presented in Table 4.2. The Table lists the 94 pairs that were examined in the first column, PARTS performance for each pair in the second, and

the results of assuming a verbal tag in the third. (The 'choice' column is explained below.) The average performance of PARTS for this task is 89%. This rate is lower than the general average performance of the tagger (95%-98% as claimed in Church 1988). The experiment (Shaked 1993), shows that assigning a verbal tag to all combinations by default will not improve but will decrease the performance rating (from 89% to 76%). This happens because in some cases the element in question is almost always a noun rather than a verb. For example, a phrasal verb like *box in* generally appears with an intervening object (*to box something in*), and thus when *box* and *in* are adjacent (except for those rare cases involving heavy NP shift) *box* is a noun. Thus we see that there is a need to distinguish between the cases where the two element sequence should be considered as one word for the purpose of assigning the lexical probability (i.e., phrasal verb) and cases where we have a noun+preposition combination, in which case PARTS' analysis will be preferred. The column labeled 'choice' in Table 4.2. shows that given a choice between the PARTS' tagging and tagging always as a verb by taking the higher performance score, the tagger will improve up to 96% for this task. In addition, some other tagging errors occur as a result of the mislabeling of phrasal verbs. For example, *to gun down* will be tagged as TOIN NN IN (preposition+noun+preposition) instead of TO VB IN (infinitive to+verb+preposition). Thus correcting the phrasal verb tagging will reduce errors in other constructions which involve phrasal verbs.

The 'choice' column in Table 4.2 reveals the need for an alternative tagging method that improves upon PARTS, but the question is, when is such an alternative needed ? An examination of the cases where PARTS had 10% or more errors shows that in most of these

cases the element in question (*gun* in *gun down*, *sum* in *sum up*, etc.) occurs much more often as a noun or an adjective. This confirms my hypothesis that PARTS will have a problem solving N/V ambiguity in cases where the lexical probability of the word points to a non-verbal element.

Table 4.2 : Performance Evaluation

PAIR	PARTS	ALL VERB	CHOICE
account-for	0.84	1.00	1.00
aim-at	0.90	0.30	0.90
amount-to	0.95	1.00	1.00
average-out	0.70	0.90	0.90
back-off	0.86	1.00	1.00
balance-out	0.92	0.84	0.92
bargain-for	0.87	0.58	0.87
bear-down	1.00	0.90	1.00
blend-in	0.91	0.95	0.95
block-in	0.97	0.02	0.97
blow-off	0.90	0.95	0.95
boil-down	0.95	1.00	1.00
book-in	1.00	0.00	1.00
bottle-up	0.36	0.90	0.90
bottom-out	0.80	0.85	0.85
box-in	1.00	0.02	1.00
break-away	1.00	1.00	1.00
break-through	0.92	0.94	0.94
bring-about	1.00	1.00	1.00
build-in	1.00	1.00	1.00
burn-down	0.95	1.00	1.00
call-back	0.96	0.84	0.96
calm-down	0.85	0.95	0.95
care-for	0.93	0.48	0.93

PAIR	PARTS	ALL VERB	CHOICE
carry-on	1.00	1.00	1.00
cash-in	0.95	0.25	0.95
chance-on	1.00	0.00	1.00
change-into	0.85	0.89	0.89
check-in	0.96	0.48	0.96
clear-out	0.87	1.00	1.00
close-down	0.77	1.00	1.00
contract-in	1.00	0.02	1.00
cool-off	0.86	1.00	1.00
cover-up	1.00	1.00	1.00
credit-with	1.00	0.00	1.00
cross-out	0.96	1.00	1.00
cross-over	0.95	0.95	0.95
cry-out	0.79	1.00	1.00
cut-back	1.00	1.00	1.00
date-from	0	1.00	1.00
deal-with	0.96	0.92	0.96
demand-of	1.00	0.04	1.00
die-out	1.00	1.00	1.00
double-up	0.80	0.95	0.95
draw-upon	1.00	1.00	1.00
dream-of	1.00	0.10	1.00
dress-up	1.00	0.90	1.00
drive-off	0.73	0.95	0.95
drop-by	0.87	0.87	0.87

PAIR	PARTS	ALL VERB	CHOICE
dry-out	0.96	0.96	0.96
ease-off	1.00	1.00	1.00
end-up	0.83	1.00	1.00
fall-in	0.92	0.29	0.92
fear-for	0.77	0.74	0.77
feed-on	0.92	1.00	1.00
feel-for	0.93	0.33	0.93
fill-in	1.00	1.00	1.00
find-out	1.00	1.00	1.00
fit-in	0.90	0.94	0.94
flesh-out	0.41	1.00	1.00
flow-from	0.94	0.42	0.94
fly-into	1.00	1.00	1.00
fool-around	0.91	1.00	1.00
force-upon	0.84	0.61	0.84
gain-on	1.00	0.09	1.00
gather-around	1.00	1.00	1.00
give-in	1.00	1.00	1.00
give-up	1.00	1.00	1.00
go-through	1.00	1.00	1.00
gun-down	0.60	0.62	0.62
hand-over	0.65	1.00	1.00
head-for	0.63	0.81	0.81
heat-up	0.94	1.00	1.00
help-out	0.95	1.00	1.00

PAIR	PARTS	ALL VERB	CHOICE
hold-down	0.92	1.00	1.00
join-in	1.00	1.00	1.00
keep-down	1.00	1.00	1.00
lead-on	1.00	0.07	1.00
let-down	0.57	0.57	0.57
lie-around	0.96	1.00	1.00
listen-in	1.00	1.00	1.00
listen-to	1.00	1.00	1.00
live-for	0.91	1.00	1.00
live-out	1.00	1.00	1.00
look-after	0.97	1.00	1.00
make-up	1.00	1.00	1.00
move-around	1.00	1.00	1.00
move-in	0.96	0.60	0.96
narrow-down	0.77	1.00	1.00
part-with	0.79	0.43	0.79
pass-along	0.96	1.00	1.00
pay-back	1.00	1.00	1.00
phone-in	0.91	0.12	0.91
pick-up	0.94	1.00	1.00
TOTAL-AVERAGE	0.89	0.79	0.96

One possible solution would be to treat these cases as one unit in the system. The lexical probability should be assigned to the pair as a whole rather than considering the two elements separately. This coincides with our representation model which requires that some pairs be classified as one unit and some as two separate words. Table 4.3 lists the cases where tagging improves 10% or more when PARTS is given the additional choice of assigning a verbal tag to the whole expression. The frequency distributions of these tokens in the Brown Corpus are presented as well, and support our earlier hypothesis of why the statistical tagger errs in these cases.

PAIR	CHOICE		BROWN FREQUENCY
date-from	1.00	date	NN/98 VB/6
flesh-out	0.59	flesh	NN/53 VB/1
bottle-up	0.55	bottle	NN/77 VB/1
hand-over	0.35	hand	NN/411 VB/8
narrow-down	0.23	narrow	JJ/61 NN/1 VB/1
close-down	0.22	close	JJ/81 NN/16 QL/1 RB/95 VB/40
drive-off	0.22	drive	NN/49 VB/46
cry-out	0.21	cry	NN/31 VB/19
average-out	0.20	average	JJ/64 NN/60 VB/6
head-for	0.18	head	JJ/4 NN/404 VB/14
end-up	0.16	end	NN/359 VB/41
account-for	0.15	account	NN/89 VB/28
double-up	0.14	double	JJ/37 NN/11 RB/4 VB/6
back-off	0.13	back	JJ/27 NN/177 RB/720 VB/26
cool-off	0.13	cool	JJ/49 NN/3 RB/1 VB/8
clear-out	0.12	clear	JJ/197 NN/1 RB/10 VB/15
calm-down	0.10	calm	JJ/22 NN/8 VB/7

Table 4.3: Samples of 10% or More Improvement

There needs to be a checking procedure that will determine which pairs will qualify/benefit from being on the all verb list. While doing that we still have to resolve cases which are really N+Prep combinations rather then V+Prt. The solution involves creating a special list for those verb-particle combinations which will benefit from an all verbs assumption and use PARTS for the rest. (insert the list in the lexicon) Then we would use simple linguistic rules to check that indeed cases of N+PP were not overlooked because of

the all verbs assumption (as a postprocessing stage to PARTS analysis).

To sum up the results, the experiment showed that PARTS, a stochastic tagger, exhibits lower then average performance in the task of identifying phrasal verbs. In addition there is no mechanism to distinguish particles from prepositions. The results show that one cause for these problems originates in the statistical approximations which consider each word in the pair separatley. It is clear that in order to tag these expressions correctly, we will have to capture additional information about the pair which is not available from the PARTS statistical model.

4.2.2.3 The Evaluation Tool: CEC

The experiment described above shows that PARTS can be improved by regarding some pairs as one unit and assigning them one lexical probability, one that will always tag them as verbs. In order to find which pairs will benefit from this option we must evaluate each pair separately. Only then can we have a list of phrasal verbs that will always benefit from this option. The list will be added to the tagger's dictionary. The corpus will be preprocessed to locate the pairs on the list and assign them a one word status so that they will be looked up in the dictionary as one item. This has to be done before the first run of the tagger. The advantage of this additional preprocessing is not only in correcting mistakes with the phrasal verbs themselves but also for neighboring elements.

The CEC, Compute, Evaluate and Compare, is a tool which automatically decides if a pair qualifies for the list of phrasal verbs that will be put in the dictionary of the tagger. The input to the program is a list of pairs to be examined and a hand-tagged corpus

consisting of 25 randomly selected sample sentences for each pair. The program will evaluate any pair given to it providing it has a small hand-tagged sample as a basis for comparison. It will then compute and evaluate the performance of PARTS and the 'all verb tag' option, on the basis of the hand-tagged sample and compare to see which of the two approaches rated better. The result of the checking procedure is a number which will indicate if and how much the tagging improved under the one verbal tag assignment. Only the cases which are improved by a verbal tag assignment will be entered on the list for potential phrasal verbs to be examined. The tool is only the first step in identifying phrasal verbs, the next steps will be explained in the next section.

Using the CEC tool one can create the same table used in the earlier experiment (Table 4.2). The 'choice' column is created by selecting the higher value of the two other columns. For example:

PAIR	PARTS	ALLVERB	CHOICE
a. hand over	65%	100%	100%-all verb
b. feel for	93%	33%	93%-PARTS
c. gun down	60%	62%	62%-all verb

Table 4.4: The CEC Table

In case **a** above the 'all verb' approach improved the tagging of the pair *hand over* from 65% to 100%, and thus it seems that the tagger will benefit from always tagging *hand over* as a phrasal verb. In contrast, case **b** shows that for the pair *feel for* PARTS' performance was much higher and thus we should rely on PARTS' tagging in this case. Relying on PARTS will cause only 7% errors. Case c is especially challenging case since it is not really clear which approach is better. The 'all verb' choice is slightly better, but either approach produce around 40% error. Thus giving a choice in those cases where the difference between PARTS' tagging the 'all verb' tagging is very small, yet the success rate of both is low, will not make much of an improvement to the system. In order to maximize the improvement, available linguistic knowledge should be used to distinguish those pairs where one verbal tag for the whole combination will be incorrect from those that might benefit from being considered one element. Thus, I am proposing a general rule stating:

18. When you see a pair which appears on the special phrasal verb list treat it as one word and assign i t one verbal tag UNLESS certain linguistic diagnostics apply.

We rely on the first run of PARTS to provide us with the contextual information needed to decide which assignment is more appropriate. In addition we use a set of pattern-matching rules that will help detect exceptions to the rule. The following section will detail the proposal for such a mechanism.

4.2.3 Enhancing the Statistical Tagger with Linguistic Information

The concern of this section is how to solve the current problem PARTS has in handling phrasal verbs. The problem originates in the statistical algorithm. The statistical model cannot capture the relation between elements in a sequence if they are separated by space. This pertains not only to the problem of phrasal verbs but to other constituents such as idioms, compounds, other multi-word verbs, etc. Thus, the treatment of phrasal verbs can shed light on these other problematic constructions and might point to a better way of handling them or at least a better understanding of the problems.

The conclusion of the previous section was that it is not enough to consider the phrasal verb as one word and assign the probability measurement to the whole unit. In some cases the tagger will still have a high error rate. Linguistic knowledge is also needed in order to handle the phrasal verb construction properly in an automatic system. I would like to explore the benefits of enhancing the statistical model with linguistic generalizations which improve the analysis. These generalizations will make use of the information obtained by the statistical analysis to resolve ambiguities that the statistical model cannot resolve.

Using the CEC tool, we are able to improve PARTS performance but not to optimize it. This is the main drawback in using the CEC tool. By making the decision to use one verbal tag for the pairs in the special list we are able to reduce the error rate but not to eliminate it. For example, in the case of *gun down*, we can improve the performance from 60% to 62% (see Table 4.4) but there will still be an error rate of 38%, which is high. By always assigning a verbal tag to the pairs in the list we lose the cases which are really a combination of Noun+Preposition as in *I put the gun down.* Thus, we want to benefit from

the choice yet not to loose points by overgeneralizing the use of this choice. Linguistic knowledge can be used to make the distinction between the cases that the rule should cover and those where the rule will lead to an error.

4.2.3.1 A Four Stage Model

We are proposing a four-stage model for processing/tagging phrasal verbs. The model improves PARTS' analysis by using the frequency counts, the results from the CEC tool, and llinguistic knowledge about the possible structural combinations in the phrasal verb construction. The model is outlined and explained in detail below.

STAGE 1: Run PARTS (current version)

STAGE 2: a. For a selected list of pairs (in the tagger's dictionary) form a one unit element tagged as a verb.

b. Run a checking procedure using linguistic diagnostics proposed below to eliminate the unit sign for the incorrect cases.

STAGE 3: Run PARTS again to ensure better tagging with new estimations.

STAGE 4: a. Open all linked pairs and assign tags to the verb and the particl.

b. Run heuristics to identify phrasal verbs which have undergone particle movemnt. Once identified, the verb and particle should be retagged to accord with the tags in 4a.

Table 4.5 shows the different processing stages according to the proposed model, for the two phrases *to gun down* and *the account for*. The two phrases require different treatments and we will show how the model can accommodate the different assignments and

correctly tag the two phrases. Using these two we will show the general operation of the model.

to gun down		**the account for**
STAGE 1:	to/TOIN gun/N down/IN	the/DET account/N for/IN
STAGE 2a:	to/TOIN gun_down	the/DET account_for
STAGE 2b:	to/TOIN gun_down	the/DET account/N for/IN
STAGE 3:	to/TO gun_down/VP	the/DET account/N for/IN
STAGE 4a:	to/TO gun/VP down/Prt	the/DET account/N for/IN
STAGE 4b:	I ordered him to gun/VB the/DET man/N down/PP I ordered him to gun/VP the/DET man/N down/Prt	

Table 4.5: A Four Stage Process

In Stage 1, the tagger PARTS (in it current mode) assigns initial tags. As a result the phrase *the account for* is correctly tagged while all components of the phrase *to gun down* are incorrectly tagged. As we discussed above, the errors follow from the fact that the tag assignment relies on frequency information and since *gun* is overwhelmingly a noun it is assigned the noun tag.

In stage 2a, using the results from the CEC tool, the model looks up the pairs that appear on our special candidates for phrasal verb list. To reiterate, the list was created in order to make sure that PARTS will benefit from a one unit labeling of the phrase. For those cases appearing on the list the general rule mentioned above in 8. and repeated here will apply:

19. When you see a pairs which appears on the special phrasal verb list treat it as one word and assign it one verbal tag UNLESS certain linguistic diagnostics apply.

"Treat it as one word" means concatenate in some way the verb and the particle so that the space between them is ignored, for example, change *gun down* to *gun_down.* This concatenation will apply to the pairs in the list every time they appear in the corpus unless there is a diagnostics in Stage 2b that applies. Both pairs *gun down* and *account for* will appear on the potential phrasal verb list.

In Stage 2b, a set of disambiguation rules will help discover the cases where a one unit labeling is incorrect and will result in an error. This checking mechanism will override the general rule whenever certain patterns are found. The pattern will make use of the information provided in the first run of the tagger (the neighboring tags) and linguistic knowledge about the possible combinations of tags (grammaticality) . The basic assumption is that we want a default VP assignment unless some more specific rule applies first. An example of such a rule is:

20. Assign a VP tag to the pair unless the tag preceding the pair is an NP starter. (An NP started can be CD, DET, ADJ, and even Noun in a compound construction.)

This one rule alone can disambiguate most of the cases where we have a real case of noun+preposition combination rather then a verb+particle/preposition case. In our case this rule only applies to one of the two phrases. In the case of *the account for* , we find a

determiner before the candidate pair thus the general concatenation rule (19) will not apply here but will apply to *to gun down*. This rule may be translated into various templates which represent the string of words which sometimes function as an NP+PP combination and sometimes as V+PART combination. We can use these templates to search for cases where the pair in question does not function as a phrasal verb and thus should not be labeled as one unit for the purpose of tagging.

In Stage 3, PARTS runs again on the corpus, which now includes the linked pairs. It will assign a verbal tag to each pair and will re-evaluate the tags for neighboring elements, which are influenced by the consideration of the pairs as one word. Thus, as the example above showed, the tag for *to* in *to gun down* will change from TOIN (prep) to TO (infinitive).

In Stage 4a, the links are eliminated, and the verb and the particle are assigned individual tags. Stage 4b applies only to nonadjacent pairs which we would like to treat in the same manner as adjacent ones. We propose to do that by using regular expression patterns This will be explained in detail in section 4.2.3.2 below.

There are several problems with this four stage proposal. In Stage 2b, we presented a rule involving NP starters. Some construction are difficult to disambiguate even using this rule and in these cases we must look for other contextual clues to help make a decision. One such case is when a verb+particle combination is preceded by a noun. Another problem involves the tag DET, which is in general a good noun/verb disambiguator. Several categories are included under DET, most of which can appear only with a noun. However, wh- determiners (such as *who, which)* can appear with both nouns and verbs. We will have to make a default decision about this determiner type. The only in the case of gerunds(VBG)

and participles(VBD) can a DET appear before a phrasal verb as in:

21a. The handing(VBG) over of the prisoner.

21b. The person who handed(VBD) over the contract.

In the case of VBD the determiner type is usually a wh-determiner which we already mentioned as problematic for N/V disambiguation. In the case of VBG, all we can say is that it is a problematic tag for PARTS in general since the tagger does not distinguish between the progressive form and the gerund. Furthermore if the gerund is considered a noun in linguistic theory then we should not consider it as a phrasal verb but a nominal derivation of it, and so it should not be tagged as a verb.

The most serious problem has to do with Stage 4a. When we open the unit and separate the pair into a verb and a particle there might be cases which are verb+preposition combinations. It is very important to be able to distinguish the two. An example is the combination *look up* which can have both a phrasal verb meaning (search) and a prepositional meaning (look upwards). In these cases we will have to rely on other types of information such as semantic knowledge, for disambiguation. The representation model described in 4.1 addresses this issue by using semantic features to distinguish the different possible senses of phrasal verbs.

To summarize, the proposed model makes use, first, of the information from statistical tagging , and second, of linguistic knowledge which does not require a sophisticated procedure yet is able to improve the taggers performance. Furthermore the identification of a disjoined particle is an important structural analysis and together Stages 2 and 4 enable a unified treatment of the phrasal verb construction whether the two elements

are adjacent or not. Stages 2 and 4 are optional and are independent of the tagger and the user can choose to ignore them. The list of phrasal verbs is flexible and more pairs could be added to it once they are checked. The procedure of testing whether a pair ought be entered in the list is automated, and it is very easy to determine if a certain phrasal verb belongs on this list. This tool can be changed to examine different categories and thus can be used to evaluate the performance of PARTS for different constructions or to test different assumptions in order to estimate PARTS' performance under various conditions.

4.2.3.2 Using Regular Expressions To Capture Particle Permutation

Particle Permutation is one of the most prominent characteristics of phrasal verbs which also causes a lot of problem to NLP systems, where it might be difficult to establish a relationship between elements that are separated. It is also difficult to capture hierarchical structure through linear representation in the case of statistical systems. With the assistance of linguistic knowledge about possible particle position and the use of regular expressions it seems possible to capture this phenomenon.

Before showing how we can use regular expressions for the disambiguation of phrasal verbs, let us define what they are and how they are being used. In computational theory a regular expression is defined as a series of characters in a form of a template which is used to facilitate the search for a string or a sequence of characters. If the search for the string using the regular expression is successful, the regular expression is said to fit the pattern. Most characters in the regular expression represent themselves. However, a few special characters called 'metacharacters' represent something other than themselves. For example:

22.

| = *or The string matches either of the two regular expressions separated by* | .

* = *Zero or more occurrences of the preceding regular expression.*

\+ = *One or more occurrences of the preceding regular expression.*

( ) = *Parentheses are used to group regular expressions so that the metacharacters [* | +] can be applied to more than one element. The parentheses take precedence over the rest of the 'metacharacters'.*

We can use regular expressions to capture phrasal verbs where the verb and the particle are not necessarily adjacent[17]. In order to do that we need to consider all the possible patterns. In the linguistic literature we find a description of possible particle positions as is discussed in detail throughout this work. When the verb and the particle are not adjacent there are a limited number of alternations that are grammatical. For our purposes the idea is to look at what can come between a verb and its particle. There are probably some border cases, but the most frequent cases fall into one of three categories:

1. A direct object = NOUN PHRASE (NP can contain several modifiers).

2. PRONOUNS .

3. A few ADVERBS (*right, quickly, slowly, all*).

No verb or preposition can occur between a verb and its particle. A test that was conducted to check several possible structures of disjoined particle showed that the NP boundary was always enough to locate the particle. In cases where there was a verb or a preposition

---

[17] Stage 4b of the four stage processing model deals with Identifying the verb and the particle when disjoined, such as in *He looked the information up.*

between the two elements, they were not related as part of the same phrasal verb. Regular expressions can be used to capture the object noun phrase which is positioned between the verb and the particle. The idea of using regular expressions to capture variety of possible combinations is also suggested by Macklovitch (1992).

Different regular expressions were tested to see which can capture the long distance dependency between the verb and its particle and the results were very encouraging. The whole template in the square brackets represents the regular expression. We were testing to see whether we can match it with existing phrasal verb patterns in the corpus without generating errors (structures that are not disjoint phrasal verbs). For example, the regular expression

[back (CD|DET)* (J|N)* N up] is made of the following components:

23.

back	*(CD|DET)**	*(J|N)**	*N*	*up*
|	|	|	|	|
verb,	*0 or more number or determiner,*	*0 or more adjective or noun,*	*noun,*	*particle*

This regular expression will select sentences where there is always one or more nouns between the verb and the particle. Using this pattern, we cannot get a [verb+pronoun+particle] combination since the pronoun is captured under the DET tag which is optional. We will never have a case of [verb+pronoun+noun+particle] structure.

If we make the N optional to capture pronoun cases we will by default generate adjacent cases which we wanted to distinguish:

24. back (CD|DET) (J|N)* up*

The following are some examples of possible usage of regular expressions to capture phrasal verbs in the corpus.[18]

Case 1: back (CD|DET)* (J|N)* N up (obligatory Noun)

1.1 The reason we have them up there is to back our troops up .

AT NN PPSS HV PPO IN RB BEZ TO VB PP NNS IN

1.2 ... made a face , and turned the magazine back side up.

... VBD AT NN , CC VBD AT NN NN NN IN

In the first example *back ...up* is tagged as a Verb ....Preposition. In the second example it is tagged as Noun ....Preposition. In both cases the tagging is correct and thus seems to violate our claim that this regular expression will capture only the correct cases of disjoint verb+particle combination. However if we look closely we see that in 1.2 *back side up* is a rare case of ready made expression.

Case 2: blow (CD|DET)* (J|N)* N up (obligatory Noun)

2.1 Stop supporting Saddam or we will blow your house up.

VB VBG NN CC PPSS MD VB PP$ NN IN '

2.2 Why didn't the minefields blow our troops up?

WRB VBD AT NNS VB PP$ NNS IN

---

[18]These samples are extracted from AP91 corpora. For each example the sentence appear in the first line and the part of speech tags for every word appear in the second line.

2.3 He just likes to blow things up.

PPS RB VBZ TO VB NNS IN

2.4 Dag apparently liked to blow things up.

NP RB VBD TO VB NNS IN

2.5 He was carrying dynamite and would blow the place up if...

PPS BEDZ VBG NN CC MD VB AT NN IN CS ...

2.6 Next he 'd be back ordering , threatening to blow people up.

JJ PPS MD BE RB NN , VBG TOIN NN NS IN

In the last case we notice the type of mistake mentioned before. The expression *to blow people up* is tagged TOIN NN NNS IN (Prep Noun Prep). Both *to* and *blow up* are mislabeled.

Case 3: break (CD|DET)* (J|N)* down (optional Noun)

Case 3 examples were extracted from the BROWN corpus, which was manually tagged and thus uses the particle tag. We expect no mistakes in labeling and we see that the regular expression was able to capture the right cases.

3.1 ...Alabama when the Freedom Riders sought to break down racial discrimination

NP WRB AT NN-TL NNS-TL VBD TO VB RP JJ NN

3.2 The insurance company or your own accounting department break down the cost of

AT NN NN CC PP$ JJ VBG NN VB RP AT NN IN

your insurance package periodically.

PP$ NN NN RB

3.3 ... as important as the more dramatic attempts to break down barriers of inequality

...QL JJ CS AT QL JJ NNS TO VB RP NNS IN NN

in the South .

IN AT NR-TL . 2 JJ

3.4 Miriam now ordered Pengally to break down the gate.

NP RB VBD NP TO VB RP AT NN

3.5 The system tended to break down during the war , but was reactivated.

AT NN VBD TO VB RP IN AT NN , CC BEDZ VBN

3.6 He felt as if he would break down and weep .

PPS VBD CS CS PPS MD VB RP CC VB

4.3 SAMPLES

This section contains several examples of how ambiguous the phrasal verb construction is and also show how the proposed model represents and correctly process difficult cases of ambiguity .

Table 4.6 below shows how one particle, "up" can combine with different verbs yielding different senses.

LITERAL	SYSTEMATIC		FIGURATIVE
GROUP 1	GROUP 2	GROUP 3	GROUP 4
stand up	come up (visibility)	write up (completion)	blow up (explode)
step up	dig up (visibility)	warm up (intensity)	brush up (freshen)
scoop up	fish up (visibility)	tie up (intensity)	call up (telephone)
run up	make up (extension)	block up(completion)	catch up (reach, join)
push up	pile up (intensity)	break up (intensity)	give up (stop)
reach up	pop up (visibility)	conjure up (visibility)	pass up (disregard)
pull up	raise up (intensity)	cut up (completion)	set up (make)
march up	stack up (intensity)	dress up (intensity)	think up (invent)
hold up	cry out (audibility)	fade up (visibility)	sign up (make a contract)

Table 4.6: The particle UP

The next example below show how one particle, "up", can belong to more than one semantic group. Also within a semantic group, particles can be associated with more than one feature:

25.

group 1: walk up "He would not just walk up to somebody and introduce himself"

group 2: offered up (extension) "They offered up a human sacrifice".

group 3: dress up (completion) "You have to dress up for this party".

hunted up (visibility) "She hunted up the references for her dissertation".

break up (intensity) "They broke up and now she is all alone".

group 4: look up "I looked the street up in the map".

stood up "He stood me up and did not even call".

It is especially hard to represent verb+particle combinations which have more than one possible sense. The examples below shows combinations with more than one meaning:

26. **break in**

26a. break in = group 4,

sense = initiate, accustom.

example = He broke in his new shoes.

26b. break in = group 1,

sense = enter in illegally

example = The house was deserted, so they broke in and took all the money.

27. **hang up**

27a. hang up = group 4,

sense = disconnect the phone.

example = He was rude so I hung up on him.

27b. hang up = group 1,

sense = put in a high place.

example = I hung the picture up on the wall.

27c. hang up = group 4,

sense = stuck.

example = I got hung up on the problem. He is hung up on this girl.

28. **break out**

28a. break out = group 1,

sense = escape.

example = He broke out of a maximum security prison.

28b. break out = group 4,

sense = appeared suddenly

example = He broke out in tears. He broke out in a grin.

28c. break out = group 4,

sense = took out

example = He broke out a bottle of wine.

According to the model, the representation of phrasal verb in the lexicon will differ depending on the type of verb+particle combination. The following examples show the

representation of literal (throw out), systematic (cry out), and figurative combinations (look up).

The examples include a description of both the verb and the particle entries of each of the combinations. In the particle entry for the literal and the systematic combinations we list all the possible grammatical structures and senses the particle can associate with the bolded ones reflect the appropriate choice in the case of the combination in question. In the figurative case (31) there will be only one entry for the whole combination.

29.

LITERAL PHRASAL VERB

	VERB ENTRY	PARTICLE ENTRY
LEXICAL REP.	[$_{v}$throw]	[$_{PART}$ out]
SYNTACTIC REP.	____(NP^{+}) PART (NP^{+}) PART = out	**VERB (NP^{+})____(NP^{+})** VERB ____(PP^{+})(THAT-S^{+})
SEMANTIC REP.	PART = out (group1)	**group 1** group 2 (audibility) .....

30.

SYSTEMATIC PHRASAL VERB

	VERB ENTRY	PARTICLE ENTRY
LEXICAL REP.	[$_{v}$cry]	[$_{PART}$ out]
SYNTACTIC REP.	____ PART(PP^{+})(THAT-S^{+}) PART = out PP = for	VERB (NP^{+})____(NP^{+}) **VERB ____(PP^{+})(THAT-S^{+})**
SEMANTIC REP.	PART = out, group 2 (audibility)	group 1 **group 2 (audibility)** .....

31.

FIGURATIVE PHRASAL VERB

	VERB + PARTICLE ENTRY
LEXICAL REP.	[$_{PV}$[$_{v}$look][$_{PART}$ up]]
SYNTACTIC REP.	____(NP^{+}) PART (NP^{+}) look (NP) up look up (NP)
SEMANTIC REP.	look up = search

For a combination such as *break down* there are three possible meanings they will be represented in the following way:

32. **break down**

32a. *The break down the building.*

[$_S$[$_{NP}$They] [$_{VP}$[$_v$broke][$_{PART}$down] [$_{NP}$the building]]]

[$_S$[$_{NP}$They] [$_{VP}$[$_v$broke] [$_{NP}$the building][$_{PART}$down]]]

Sense: The combination belongs to group1 thus the meaning is compositional from the individual meaning of the verb and the particle. The phrasal verb will have a directional sense.

32b. *The car broke down on the street.*

[$_S$[$_{NP}$The car][$_{VP}$[$_v$broke][$_{PART}$down] [$_{PP}$on the street]]]

Sense: The combination belongs to group 3 and the particle is associated with the feature [intensity]. The verb in this type combination retains its basic meaning and the particle meaning is non-directional. The meaning of *break* is intensified by the particle *down* in the sense of *become disabled.*

32c. *He broke down in tears.*

[$_S$[$_{NP}$He][$_{VP}$[$_{PV}$[$_v$broke][$_{PART}$down]] [$_{PP}$in tears]]]

Sense: *collapsed*

The meaning of the composition is non compositional and cannot be predicted from the sense of its components. Thus it should be listed in the lexicon.

CHAPTER V: CONCLUSION

5.1 SUMMARY AND CONCLUSIONS

The major contribution of this work is in highlighting the problem of representing and processing phrasal verbs in both theoretical and computational linguistics.

Chapter I makes the claim that there is no one definition in the literature that covers all types of phrasal verbs, and provides a definition for the purposes of this research. Chapter II shows that phrasal verb ambiguity can create serious problems for NLP systems, and illustrates how mishandling them can lead to other processing errors. We cover three main NLP areas: tagging, structural analysis, and semantic classification. Chapter III lays out the linguistic description of the construction, showing that theoretical linguistics cannot offer a solution to the problems described in Chapter II, because there is no single, coherent account of the structure and semantics of the verb+particle combinations. In addition, while theoretical linguistics can rely on the speakers' intuition to resolve ambiguity while computational linguistics must provide an accurate representation to deal with such ambiguity. Chapter IV proposes solutions for the problems presented in Chapter II that deal with representing phrasal verbs in the computational lexicon. We present a three-level representation model which classifies phrasal verbs into three groups: literal, systematic, and figurative. Each class is treated differently to accommodate its unique characteristics and

enable accurate processing and interpretation.

In the computational linguistics literature we find no account of the different types of phrasal verbs and no attempt to deal with their treatment in the lexicon. Phrasal verbs are usually treated as idiosyncratic items in a special list. Our model is unique in that it offers broad coverage and systematic representation of phrasal verbs by classifying them into semantic classes and treating each class in a slightly different way under the framework of the general model. The basic assumption is that in order to accurately represent phrasal verbs, the relationship between the verb and the particle must be considered. In contrast to other frameworks (such as COMLEX), the particle is assigned its own lexical entry where information about all possible grammatical structures and meanings is listed. In this way the model accounts for the fact that the particle can combine with different verbs yielding different senses, and at the same time accounts for the fact that the same verb can interact with more than one particle.

At the lexical level, the particle is assigned its own part of speech tag (PART) and is not collapsed under either the preposition or adverbial tags. The verb and the particle in the literal and systematic combinations are listed separately under regular lexical entries, while the figurative combinations are put in a special idiom list. The pairs in this list are regarded as one unit and are assigned one tag (PV) in addition to the regular internal tagging (VERB+PART).

At the syntactic level the particle entry lists all possible grammatical structures in which the particle can appear. The verb entry lists the specific particles that can be associated with the verb. The model advocates collapsing the subcategorization frames in a way that

will account for the permissible structures in the most parsimonious format. To do that the model uses special notation to indicate the fact that some elements in the subcategorization frames cannot appear together in one clause.

At the semantic level, the model uses semantic groups and features to differentiate between possible semantic classes of phrasal verb combinations. In the particle entry all the possible senses are listed, coupled with the group number and features associated with them (as described in detail in section 4.1.3). By using features we can account for the multiple meanings the particle can have in association with different verbs and also account for the fact that the same verb+particle combination can have more than one meaning. This is possible if the features associated with the particle are different for each sense. In these cases, the particle is listed in the verb entry more than once.

The model presented in this work needs to be tested. It might turn out that its main advantage is in the classification and characterization of phrasal verb and that it needs to be refined in order to work efficiently in an NLP system. This could be especially true for example for a machine translation system which requires a very specific sense translation. Translating phrasal verbs between languages (especially those which do not have such a construction at all) is a very demanding task.

A partial implementation of these solutions suggested in the first part of Chapter IV are presented in the second part. The results of several experiments designed to test the implementation point to the advantages of using a grammar-based approach coupled with statistical approximation to resolve language ambiguity.

5.2 FUTURE RESEARCH

This research deals with problems of handling phrasal verbs in both a theoretical and a computational framework. Other problematic constructions which display similar syntactic and semantic behavior should be considered too. Constructions such as compounds and multi-word verbs are similar in their syntactic behavior, while idioms are similar in their semantic respects. In computational linguistics today there are no systematic proposals for handling these elements, yet they generally recognized to be problematic. This research proposes a model for phrasal verb representation. Future work should explore the possibility of expanding the model to account for other similar constructions, such as those mentioned above.

Another problem that is addressed in our treatment of phrasal verbs is the identification and labeling of disjoint elements, in this case verbs and particles. This touches on the issue of long-distance dependencies, which are especially difficult for statistical systems to handle, because they rely mostly on linear processing and use very limited lookahead to mark decisions. Thus, special attention should be given in the future to finding a way to capture different cases of long distance dependencies along the lines suggested in my model. The tests reported in Chapter IV of a partial implementation of the model dealt only with tagging issues. The CEC tool has so far been used to discover which phrasal verbs are problematic for the statistical tagger PARTS. It could easily be turned into a checking mechanism for any construction that requires evaluation. Efforts should also be put into implementing the structural and the semantic levels of the model to test and determine to what extent they improve performance.

Finally, this work advocates the use of both linguistic and statistical knowledge to resolve ambiguity. We show how the combination of the two frameworks can facilitate phrasal verb processing. Future work should explore other areas where the combination of the two approaches can improve the performance of NLP systems.

BIBLIOGRAPHY

Aarts, B. (1989) Verb Preposition and Small Clauses in English. *Linguistics.* Vol. 25 No.2: 277-290.

Aronoff, M. (1976) *Word Formation in Generative Grammar.* Mass.: MIT Press.

Bobrow, S. and S. Bell (1973) On catching on to Idiomatic Expressions. *Memory and Cognition.* No.1: 343-346.

Bolinger, D. (1971) *The Phrasal Verb in English.* Cambridge, Mass.: Harvard University Press.

Boguraev, B. and J. Pustejovsky (1993) Lexical Ambiguity and The Role of Knowledge Representation in Lexicon Design. *Artificial Intelligence.* Vol 63. pp: 193-23.

Brannon, L. (1975) On The Understanding of idiomatic Expressions. Unpublished Dissertation. University of Texas.

Chomsky, N. (1957) *Syntactic Structures.* New York: Mouton.

Chomsky, N. (1970) Remarks on Nominalization. in R. Jacobs and P Rosenbaum (eds.), *Readings in English Transformational Grammar.* The Hague: Mouton. pp: 84-221.

Chomsky, N. (1972) *Studies On Semantic in Generative Grammar.* The Hague: Mouton.

Church, K. W. (1988) A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing,* Austin, Texas, 1988, pp: 136-143.

Church, K. W., W. Gale, P. Hanks and D. Hindle (1991) Using Statistics in Lexical Analysis. In Uri Zernik (ed.), *Exploring Online Resources to Build a Lexicon.* Hillsdale: Lawrence Erlbaum Press.

Church, K. W. and P. Hanks (1990) Word Associations Norms, Mutual Information and Lexicography. *Computational Linguistics.* Vol. 16. No. 1.

Church, K. W. and R. Mercer (1993) Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics.* Vol.19, No. 1

Clark H. and P. Lucy. (1975) Understanding What is Meant from What is Said: a Study in Conversationally Conveyed Requests. *Journal of Verbal Learning and Verbal Behavior.* No. 14: 56-72.

Dagut, M. and B. Laufer (1985) Avoidance of Phrasal Verbs: A Case of Contrastive Analysis. *Studies in Second Language Acquisition.* No. 7: 73-79.

den Dikken, D. (1991) *Particles.* The Netherlands: Holland Institute of Generative Linguistics.

De Rose, S. J. (1988) Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics.* No. 14: 31-39.

Dowty. D. R. (1979) *Word Meaning and Montague Grammar.* Dordrecht: Kluwer Academic Publishers.

Emonds, J. (1976) *A Transformational Approach to English Syntax.* New York: Academic Press.

Foster, G. F. (1991) *Statistical Lexical Disambiguation.* Masters Thesis, McGill University, School of Computer Science. Montreal, Canada.

Fodor, J. A. and J. J. Katz. (1964) *The Structure of Language: Readings in the Philosophy of Language.* New Jersey: Prentice Hall.

Francis, W. and H. Kucera. (1982) *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin.

Fraser, B. (1965) *The Verb Particle Combination in English.* PhD. Dissertation. MIT.

Fraser, B. (1976) *The Verb Particle Combination in English.* New York: Academic Press.

Gibbs, R. W. (1980) Spitting The Beans on Understanding and Memory for Idioms on Conversation. *Memory and Cognition.* No. 8: 149-159.

Gibbs, R. W. (1986) Skating on Thin Ice: Literal Meaning and Understanding Idioms in Conversation. *Discourse Processes.* No. 9: 17-30.

Grice, H. P. (1975) Logic and Conversation. In P. Cole and J. T. Morgan (eds.), *Syntax and Semantics: Speech Act 3.* New York: Academic Press.

Harris, Z. S. (1951) *Structural Linguistics.* Chicago: The University of Chicago Press.

Harris, Z. S. (1970) *Papers in Structural and Transformational Linguistics.* Dordrecht: Reidel.

Harris, Z. S. (1988) *Language and Information.* New York: Columbia University Press.

Hindle, D. (1990) Disambiguating TO. Unpublished manuscript. AT&T Bell Labs.

Hirschberg, J. (1992) Pitch Accent in Context: Predicting Intonational Prominence from Text. *Artificial Intelligence:* No. 63: 305-340.

Irujo, S. (1986) Don't Put Your Leg in Your Mouth: Transfer in the acquisition of Idioms in Second Language. *TESOL Quarterly.* Vol. 20 No. 2: 287-304.

Katz, J. and P. Postal. (1963) Semantic Interpretation of Idioms and Sentences Containing Them. *Quarterly Progress Report, MIT Research Lab. of Electronics 70.* pp: 275-282.

Kayne, R. S. (1985) Principles of Particle Constructions. In J. Gueron, H. G. Obenauer and J. Y. Pollock. (eds.) *Grammatical Representation.* Dordrecht: Foris.

Kayne, R. S. (1993) The Antisymmetry of Syntax. CUNY, Graduate School. Unpublished manuscript.

Kennedy, A. G. (1920) The Modern English Verb Adverb Combination. *Stanford University Publications in Language & Literature.* Vol.1 No.1. Stanford, CA.

Kochan, S. D. and P. H. Wood (1991) *UNIX Shell Programming.* Carmel, Indiana: Hayden. pp: 477-482

Kroch, A. S. (1979) Review of Fraser, B. (1976) The Verb Particle Combination in English. *Language 55.* Vol.1: 219-224.

Le Roux, C. (1988) On the Interface of Morphology and Syntax. Evidence from Verb-Particle Combinations in Afrikaans. *SPIL* No.18. November 1988. M.A. Thesis. University of Stellenbosch.

Liberman, M. Y. (1991) The Trend towards Statistical Models in Natural Language Processing. in E. Klein and F. Veltman (eds.), *Natural Language and Speech Symposium Proceedings.* Brussels: Esprits.

Lieber, R. (1984) *On the Organization of the Lexicon.* Ph.D. Dissertation, University of New Hampshire.

Lipka, L. (1972) *Semantic Structure and Word-Formation Verb-Particle Constructions in Contemporary English.* Munchen: Wilhelm Fink Verlag.

Macklovitch, E. (1992) Where the Tagger Falters. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.* pp: 113-126.

Macleod, C. and R. Grishman (1994) COMLEX Syntax Reference Manual (Ver. 3.0) New York University. Unpublished manuscript.

Macpharson, M. (1991) Redefining the 'Level' of The 'Word'. In J. Pustejovsky and S. Bergler (eds.), *Proceeding of the 1st SIGLEX workshop.* CA.

McGilton, H. and R. Morgan (1983) *Introduction to the UNIX System.* New York: McGroaw Hill . pp: 152-156.

Marcus, M., B. Santorini and D. Magerman. (1992) First Steps Towards An Annotated Database of American English. Dept. of Computer and Information Science, University of Pennsylvania, MS.

McPartland, P. (1983) What the Non-native Speakers Leave Out. Unpublished Manuscript. CUNY, Graduate School.

McPartland, P. (1989) *The Processing of Phrasal Verbs by Native and Non-Native Speakers of English.* Ph.D. Dissertation. CUNY, Graduate School.

Meeter, M., R. Schwartz and R. Weischedel (1991) POST: Using Probabilities in Language. *Proceedings of JACAI 91.* Sydney, Australia.

Milne, R. (1983) *Resolving Lexical Ambiguity in a Deterministic Parser.* PhD. Dissertation, University of Edinburgh.

Milne, R. (1986) Resolving Lexical Ambiguity. In Uri Zernik (ed.), *Exploring Online Resources to Build a Lexicon.* Hillsdale: Lawrence Erlbaum Press.

Mudler, R. (1992) Datives and Particle Alternation. In S. Barbiers, M. den Dikken and C. Levelt (eds.), *Proceedings of the Third Leiden Conference for Junior Linguistics.* Leiden: Leiden.

Nilsen, D. (1972) *English Adverbials.* The Hague: Mouton.

Palmer, F. R. (1974) *The English Verb.* London: Longman.

Pelli, M. G. (1976) Verb Particle Constructions in American English. Zurich: Francke Verlag Bern.

Radford, A. (1981) *Transformational Syntax.* Cambridge: Cambridge University Press.

Ravin, Y. (1989) Disambiguating and Interpreting Verb Definitions. *Proceedings of 28th Annual Meeting of the Association of Computational Linguistics.* pp: 260-267.

Ross, K. (1992) *Modeling of Intonation for Speech Synthesis.* PhD Dissertation Proposal. Boston University.

Ross, K. M. Ostendorf and S. Shattuch Hufnagel. (1992) Factors Affecting Pitch Accent Placement. In *Proceeding of ICSLP.* Banff, Canada. October 1992.

Searle, J. R. (1969) *Speech Acts.* London: Cambridge University Press.

Selkirk, E. O. (1982) *The Syntax of Words.* Cambridge Mass.: MIT Press.

Shaked, N. A. (1993) How Do We Count? The Problem of Tagging phrasal Verbs in PARTS. In *Proceedings of 31st Annual Meeting of the Association of Computational Linguistics.* pp: 289-291.

Shannon, C. and W. Weaver. (1964) *The Mathematical Theory of Communication.* Urbana: The University of Illinois Press.

Simpson, J. (1983) Discontinuous Verbs and the Interaction of Morphology and Syntax. In *Proceedings of the West Coast Conference on Formal Linguistics 2.* Stanford, CA.

Smadja, F. (1991a) *Extracting Collocations from Text. An Application: Language Generation.* Ph.D. Dissertation. Columbia University.

Smadja, F. (1991b) Macrocoding the Lexical with Co-occurrence Knowledge. In Uri Zernik (ed.), *Exploring Online Resources to Build a Lexicon.* Hillsdale: Lawrence Erlbaum Press.

Smith, S. P. (1925) *Words and Idioms: Studies in the English Language.* London: Constable and Co.

Sroka, K. A. (1972) *The Syntax of English Phrasal Verbs.* The Hague: Mouton.

Stowell, T. (1981) *Principles of Lexical and Phrasal Structure.* Ph.D. Dissertation. MIT.

Swinney, D.A. (1982) The Structure and Time Course of Information Interaction During Speech Comprehension: Segmentation, Access, and Interpretation. In J. Mehler, E. Walker and M. Garrett (eds.), *Perspectives on Mental Representation.* Hillsdale, New Jersey: L. Erlbaum Associates.

Swinney, D.A. and A. Culter (1979) The Access and Processing of Idiomatic Expressions. *Journal of Verbal Learning and Verbal Behavior.* No. 18: 523-534.

Vestergoard, T. (1977) *Prepositional Phrases and Prepositional Verbs: A Study in Grammatical Function.* The Hague: Mouton.

Weinreich, V. (1979) Problems in The Analysis of Idioms. In J. Puhvel (ed.), *Substance and Structure of Language.* Berkeley: University of California Press. pp: 23-81.

Yorio, C. A. (1989) Idiomaticity as an Indicator of System Proficiency. In K. Hyltenstam and L. Obler (eds.), *Bilingualism across the Life Span: Aspects of Acquisition, Maturity and Loss.* New York: Cambridge University Press.

AUTOBIOGRAPHICAL STATEMENT

Nava Ayala Shaked was born on August 30, 1963 in Ramat-Gan, Israel. She received a B.A. degree in English Linguistics and Literature from Bar-Ilan University in 1987. In September 1988 she enrolled in the Linguistics Department in CUNY Graduate School, New York, where she is expected to receive the Ph.D. in Linguistics in September 1994.

From January 1991 to January 1993 Nava worked in the Linguistic Research group at AT&T Bell Labs. From February 1993 to September 1994 she worked in the Speech Recognition and Language Understanding group in NYNEX Science & Technology. Her work in both research institutions covered different aspects of speech synthesis and recognition.

Ms. Shaked is a member of the Association of Computational Linguistics (ACL).