Verbal Polysemy Resolution through Contextualized Clustering of Arguments

A Dissertation

Presented to

The Faculty of the Graduate School of Arts and Sciences

Brandeis University

Department of Computer Science

James Pustejovsky, Dept. of Computer Science, Advisor

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Anna A. Rumshisky

February, 2009

The signed version of this signature page is on file at the Graduate School of Arts and Sciences at Brandeis University.

This dissertation, directed and approved by Anna A. Rumshisky's committee, has been accepted and approved by the Graduate Faculty of Brandeis University in partial fulfillment of the requirements for the degree of:

DOCTOR OF PHILOSOPHY

Adam Jaffe, Dean of Arts and Sciences

Dissertation Committee:
James Pustejovsky, Dept. of Computer Science, Chair
Martin Cohn, Dept. of Computer Science
Pengyu Hong, Dept. of Computer Science
Patrick Hanks, Faculty of Informatics, Masaryk University
Dekang Lin, Google Inc.

©Copyright by

Anna A. Rumshisky

2009

Acknowledgements

I would like to thank my advisor, James Pustejovsky, for creating a very nurturing and creative work atmosphere, as well as supporting and encouraging the spirit of productive collaboration. During the time that I have been privileged to work with him, his intellectual vision and the ability to lead has earned my sincere admiration. I am immensely grateful to Patrick Hanks for his support and continuous inspiration, especially with respect to working with linguistic data. Without our discussions, the ideas in this dissertation would not have been formulated. An important influence has also been my internship time at Google, which provided me with an opportunity to think about the data in a variety of new ways.

I would like to thank the members of my dissertation committee, Dekang Lin, Marty Cohn, and Pengyu Hong, for their valuable comments during different stages of my dissertation work from the initial conception to the final revisions. I am indebted to the colleagues in the Faculty of Informatics in Masaryk University, particularly to Karel Pala and Pavel Rychly, and also to the colleagues in the Berlin-Brandenburgische Akademie der Wissenschaften in Berlin for their comments on some of the earlier stages of this research. I would also like to thank Alessandro Lenci and the other reviewers who provided suggestions and comments on the publication of parts of this work in the Italian Journal of Linguistics. I am also very grateful to

Olga Batiukova for her support and help with the final stages of data annotation, including both data analysis and adjudication.

I would also like to acknowledge the immense encouragement I received from my family while completing this work. I would like to thank my mom for her unending support and my dad for his tireless contribution to sustaining my motivation. I am extremely grateful to my brother for the numerous inspiring conversations and his invaluable participation. I am certain that the chalk marks we left on the streets of Watertown while discussing various ideas were a source of constant entertainment to the local inhabitants.

My thanks also go to all my friends who have patiently listened to me while I formulated my ideas and who supported me all throughout this process. I would particularly like to acknowledge Sergey Bratus for his continuous subversive influence on my academic choices and Antonella Di Lillo for helping me stay in good spirits during the late-night writing sessions. I would also like to thank all my colleagues and friends at the Brandeis Computer Science Department for their support and encouragement.

Abstract

Verbal Polysemy Resolution through Contextualized Clustering of Arguments

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

by Anna A. Rumshisky

Natural language is characterized by a high degree of polysemy, and the majority of content words accept multiple interpretations. However, this does not significantly complicate natural language understanding. Native speakers rely on context to assign the correct sense to each word in an utterance. NLP applications, such as automated word sense disambiguation, require the ability to identify correctly context elements that activate each sense.

Our goal in this work is to address the problem of contrasting semantics of the arguments as the source of meaning differentiation for the predicate. We investigate different factors that influence the way sense differentiation for predicates is accomplished in composition and develop a method for identifying semantically diverse arguments that activate the same sense of a polysemous predicate. The method targets specifically polysemous verbs, with an easy extension to other polysemous words. The proposed unsupervised learning method is completely automatic and relies exclusively on distributional information, intentionally eschewing the use of human-constructed knowledge sources and annotated data. We develop the notion of selectional equivalence for polysemous predicates and propose a method for contextualizing the representation of a lexical item with respect to the particular context

provided by the predicate.

We also present the first attempt at developing a sense-annotated data set that targets sense distinctions dependent predominantly on semantics of a single argument as the source of disambiguation for the predicate. We analyze the difficulties involved in doing semantic annotation for such task. We examine different types of relations within sense inventories and give a qualitative analysis of the effects they have on decisions made by the annotators, as well as annotator error.

The developed data set is used to evaluate the quality of the proposed clustering method. The output is adapted for evaluation within a standard sense induction paradigm. We use several evaluation measures to assess different aspects of the algorithm's performance. Relative to the baselines, we outperform the best systems in the recent SEMEVAL sense induction task (Agirre et al., 2007) on two out of three measures.

We also discuss further extensions and possible uses for the proposed automatic algorithm, including the identification of selectional behavior of complex nominals (Pustejovsky, 1995) and the disambiguation of noun phrases with semantically weak head nouns.

Contents

Abstract **vi**

1 Introduction **1**
1.1 Problem of Polysemy Resolution 1
1.2 Focus of Work 6
1.3 Approach 7
1.3.1 Clustering Algorithm 7
1.3.2 Sense-Annotated Data Set 8
1.4 Practical Applications 9
1.5 Outline 10

2 Related Work **12**
2.1 Defining a Sense Inventory 12
2.1.1 Lexicographic work 13
2.1.2 Theoretical work 16
2.2 Distributional Models for Semantic Similarity 23
2.2.1 Context representation 24
2.2.2 Similarity measures 26
2.2.3 Proposals for sense detection 27
2.2.4 Evaluation 37

3 Resolving Polysemy in Context **44**
3.1 Selection and Compositionality 44
3.1.1 Reusability of semantic features 46
3.1.2 Selectors and sense separation 47
3.2 Problems with Sense Inventories 49
3.2.1 Defining sense categories 49
3.2.2 Boundary cases 49
3.2.3 Regular semantic processes 50
3.2.4 Parallel sense distinctions 52
3.2.5 Semantic underspecification 53

3.3 Summary 53

4 Bipartite Contextualized Clustering 54
4.1 Preliminaries 54
4.1.1 Motivation 54
4.1.2 Contextualized Similarity 56
4.1.3 Selectional Equivalence 57
4.2 System Architecture 59
4.2.1 Algorithm Description 60
4.2.2 System Configurations 64
4.2.3 Implementation 69
4.3 Summary 71

5 Argument-based Sense Annotation 72
5.1 Motivation 72
5.2 Task Description 73
5.2.1 Data set construction 73
5.2.2 Defining the task for the annotators 75
5.3 Annotation Interface 75
5.4 Systematic Relations Between Senses 79
5.4.1 Argument structure alternations 79
5.4.2 Event structure modification 80
5.4.3 Lexical semantic features 80
5.4.4 Metaphor and metonymy 81
5.5 Analysis of Annotation Decisions 82
5.6 Summary 86

6 Evaluation via Word Sense Induction 87
6.1 Motivation 87
6.2 WSI Algorithm 88
6.2.1 Cluster rank 88
6.2.2 Selector-cluster association 88
6.2.3 Using clusters in a WSI task 89
6.3 Data Set 90
6.4 Evaluation 90
6.5 Summary 94

7 Computational and Theoretical Extensions 96
7.1 Sense Selection in Dot Nominals 96
7.1.1 Data Analysis for Dot Objects 96
7.1.2 Clustering Task 101

7.2 Modifier-Based Disambiguation of NPs 106
7.3 Summary 108

8 Conclusions **113**

Appendices **117**

A Resources **118**
A.1 Corpora, Parsers, and Lexical Resources 118
A.2 The Sketch Engine 119

B Annotation Guidelines **123**
B.1 General Instructions 123
B.2 Verb-Specific Instructions and Sense Inventories 124

C Test Data **132**
C.1 Verbs 132

Bibliography **156**

List of Tables

2.1 CPA pattern grammar 42
2.2 Similarity measures 43

4.1 Licensing contexts for selectors of *take on* 58
4.2 System configurations 65
4.3 Selectors for *deny* 67
4.4 Similarity computation for selectional equivalents of *deny* 68
4.5 Similarity matrix for selectional equivalents of *deny* 69

6.1 Per-word characteristics of the data set and system performance . . . 91
6.2 Performance of our system for different clustering configurations . . . 94
6.3 SEMEVAL Task-2 system performance 94

7.1 Inventory of Complex Types 109
7.2 Selectors for *lunch* 110
7.3 Similarity computation for contextual synonyms of *lunch* 110
7.4 Dendrogram trace for the target *lunch*, seed *conference*. 111
7.5 Selector assignment for *lunch* as direct object 112

List of Figures

4.1 Merging ranked selector lists 64
4.2 Selectors for *deny* 67
4.3 Processing Flow 70

5.1 Annotation interface: Target selection 76
5.2 Annotation interface: Predicate sense disambiguation for *deny* 77
5.3 Annotation interface: Instructions display 78
5.4 Adjudication interface 78

7.1 Choosing selectors for the noun pair *lunch-n/conference-n* 103
7.2 Intra-cluster APS for *lunch* 105

Chapter 1

Introduction

It is a well-known phenomenon that within a natural language, the same word often has multiple interpretations. This thesis is concerned with the automatic resolution of such ambiguities, particularly as applied to the domain of verbal polysemy. There are a number of complex factors that allow fluent speakers to identify the appropriate sense of a polysemous word in context. In this thesis, we will focus on one of the least studied factors, namely, the contribution of the semantics of the arguments towards differentiating the senses of polysemous verbs.

1.1 Problem of Polysemy Resolution

From a linguistic perspective, it is common to assume that the meaning we associate with an utterance is constructed compositionally. Namely, that the expression we relate to a sentence is built out of the meanings that we associate with its component parts. This is standardly called the Fregean "principle of compositionality" and has guided much of the semantic work in linguistics for the last 50 years. This view of compositionality, while satisfying the criteria of the idealized model, runs into significant problems when looking at real data. It proves to be problematic for several reasons. First, the component parts entering into deriving the complex expression may be and often are themselves ambiguous. Secondly, the meaning of this complex expression will depend for its full interpretation on the larger context in which it is embedded.

For humans, a high degree of polysemy in language does not appear to significantly complicate our understanding process. This is due to the fact that within a specific context, each word is usually assigned a single interpretation, and the polysemy of an expression is significantly reduced or eliminated completely. In lexical semantics, two kinds of lexical ambiguity are traditionally distinguished: polysemy and homonymy. *Polysemy* is the ambiguity between related meanings of the same lexical form, while

homonymy is the term used for the ambiguity between completely unconnected, often diachronically distinct meanings. *Regular polysemy* is further distinguished as the kind of polysemy in which the word's meanings are related in a regular and predictable way, and the same logical relationship between meanings characterizes other polysemous words in the language.

The phenomena of regular polysemy, as well as polysemy in general, are exhibited by all the major word classes. The meaning assigned to the word is determined by a combination of contextual factors relevant for that particular word class. For example, the multiple interpretations that can be carried by the adjectives such as *hot* and *fast* in (1.1) are eliminated in each usage context, and the specific meaning assigned to each adjective is effectively a function of the semantics of the head noun.

(1.1) a. hot chocolate
b. hot football player
c. fast typist
d. fast woman

In case of regular polysemy in nouns, the governing verb or a modifier often determines which meaning is assigned to the noun, as can be seen in (1.2) for the noun *newspaper*.

(1.2) a. For this game, one needs a hard-boiled egg and a *rolled-up newspaper*. (*physical object*)
b. The government responded to reports in *conservative newspapers*. (*organization*)
c. The boy who *delivers* the *newspaper* came by. (*physical object*)
d. They *accused* the *newspaper* of making up stories about them. (*organization*)

Two contexts of occurrence of the same word may be quite similar and yet activate different meaning for a given word. Consider the interpretation of the word *newspaper* below.

(1.3) a. I started this newspaper two years ago.
b. I finished this newspaper two hours ago.

Generative Lexicon (GL) (Pustejovsky, 1995) offers a system of compositional mechanisms that account for such variation of sense in context. In particular, complex types are introduced to deal with regular polysemy of such nouns, and the mechanisms of *coercion* are proposed to account for type selection, exploitation, or shifting

(cf. Pustejovsky, 1995; Pustejovsky, 2006). A set of *qualia roles* associated in GL with each noun (specifying, for example, purpose or origin of the object) help to account for the polysemy of its adjectival modifiers, as well as for sense alternations of the noun itself. Corpus Pattern Analysis (CPA) (Pustejovsky et al., 2004; Hanks and Pustejovsky, 2005; Rumshisky et al., 2006) is a lexicographic word analysis technique that aims to record contextual cues that typically distinguish between different senses of each word, using an extended set of context features for this task. We take the inspiration from Generative Lexicon and Corpus Pattern Analysis in our analysis of the mechanisms involved in sense disambiguation and in constructing the computational model to handle sense selection phenomena.[1]

Within the scope of a sentence, the meaning that gets assigned to a word is usually determined by a combination of two factors: (1) the syntactic frame into which the word is embedded, and (2) the semantics of the words with which it forms syntactic dependencies. We will use the term *selector* to refer to such words, regardless of whether the target word is the headword or the dependent element in the syntactic relation. In this work, our primary focus is the resolution of polysemy in verbs. The term "syntactic frame" above should be understood broadly as extending to minor categories (such as adverbials, locatives, temporal adjuncts, etc.) and subphrasal cues (genitives, partitives, negatives, bare plural/determiner distinction, infinitivals, etc.).

The set of all *usage contexts* in which a polysemous word occurs can usually be split into groups where each group roughly corresponds to a distinct *sense*. Consider a subset of senses of the verb *absorb* in (1.4), where three very distinct meanings of the verb are represented[2].

(1.4) a. The customer will *absorb* the cost.
Mr. Clinton wanted energy producers to *absorb* the tax.
(*pay; take on an expense*)

b. They quietly *absorbed* this new information.
Meanwhile, I *absorbed* a fair amount of management skills.
(*learn*)

c. The villagers were far too *absorbed* in their own affairs.
He became completely *absorbed* in struggling for survival.
(*preoccupy*)

In each case, certain specific context element(s) activate the appropriate sense of the target word. Given the data above, the prototypical norms of usage for the verb

[1] CPA and GL are reviewed in more detail in Chapter 2.

[2] These and other examples here are taken, in somewhat modified form, from the British National Corpus (BNC).

absorb are recorded in CPA in terms of the following context patterns:[3]

(1.5) a. [[Person]] absorb [[LEXSET Asset: tax, cost, ...]]
b. [[Person]] absorb {([QUANT]) [[Information]]}
c. [[Person]] {be | become} absorbed {in [[Activity]] | [[Abstract]]}

While syntactic frame clearly contributes to the activation of the third sense, argument semantics is what distinguishes between the first two senses.

As the appropriate context elements are added into consideration, each of them is locked in a *pattern sense*, i.e. the sense it acquires within this particular context, as each element's "meaning potential" (Halliday, 1973) is realized. Elements of the pattern carry different weight in disambiguating a given polysemous word. One can think about this in terms of adjusting the probability distribution on senses. Consider the verb *fire* which, for the sake of the argument, we will assume to have only three main senses: (1) *shoot* (as in, "fire bullets, rounds, shots; fire guns and guns firing at targets and on people and human groups"), (2) *dismiss from a job* ("fire from a job"), and (3) *inspire* ("fire enthusiasm, spirit, interest, imagination"). Without any information about the context, we assume a certain prior distribution $P(sense_i) \sim (p_1, p_2, p_3)$. Now we add the information about a particular semantic feature of the subject (e.g. *Animate*). Perhaps, that doesn't do very much to distinguish between the first two senses, but it makes the third sense much less likely. So given the context element c_1 = (subject, Animate), an adjusted distribution might be $P(sense_i|c_1) \sim (p'_1, p'_2, 0)$. Next, perhaps, we add the information about some semantic feature of direct object. For example, that it denotes a *Firearm* (i.e. the next context element c_2 = (object, Firearm)). This most likely will eliminate all senses but one, giving the distribution $P(sense_i|c_1, c_2) \sim (1, 0, 0)$.

In some cases, a more extended context is required to resolve the indeterminacy. For example, consider the two senses of the verb *watch* in (1.9):

(1.6) a. Sense 1: *to see and attend; to follow while looking*
We *watched* the train pull into the station.
He was *watching* the circling helicopters.

b. Sense 2: *to know and attend intellectually; to follow by being aware*
The French government had to *watch* the first partition of Poland in 1772.
I have *watched* his career develop and know he is ready for the challenge.

Distinguishing between these two cases is often impossible without additional context,

[3]Double square brackets are used for argument type specification, curly brackets are used for syntactic constituents, and parentheses indicate optionality. For full pattern syntax, see Chapter 2.

for example, as in (1.7).

(1.7) She *watched* her group's progress with interest.

But typically a clause or a sentence context is sufficient for the disambiguation, with the syntactic frame and/or one or more arguments or adjuncts contributing to sense assignment.

To illustrate the contribution of different context parameters to disambiguation, consider the verbs in (1.8) and (1.9). Syntactic patterns for the verb *deny* in (1.8) disambiguate between the two dominant senses: (i) *proclaim false* and (ii) *refuse to grant*.[4]

(1.8) Syntactic frame:

a. The authorities *denied* that there is an alternative. [that-CLAUSE]
The authorities *denied* these charges. [NP]
(*proclaim false*)

b. The authorities *denied* the Prime Minister the visa. [NP] [NP]
The authorities *denied* the visa to the Prime Minister. [NP] [to-PP]
(*refuse to grant*)

For the senses of *fire*, *absorb*, *treat*, and *explain* shown in (1.9), contrasting argument and/or adjunct semantics is the sole source of meaning differentiation. The relevant argument type is shown in brackets and the corresponding sense in parentheses:

(1.9) Semantics of the arguments and adjuncts/adverbials:

a. The general *fired* four *lieutenant-colonels*. [PERSON] (*dismiss*)
The general *fired* four *rounds*. [PHYSOBJ] (*shoot*)

b. The customer will *absorb* this *cost*. [ASSET] (*pay*)
The customer will *absorb* this *information*. [INFORMATION] (*learn*)

c. This new *booklet explains* our strategy. [INFORMATION] (*describe, clarify*)
This new *development explains* our strategy. [EVENT] (*be the reason for*)

d. Peter *treated* Mary with *antibiotics*. [with MEDICATION] (*medical*)
Peter *treated* Mary with *respect*. [with QUALITY] (*human relations*)

[4]These and other examples are taken, in somewhat modified form, from the British National Corpus (BNC, 2000).

The appropriate sense in the examples above is identified by looking solely at the semantics of the direct object, subject, or adjunct, respectively.

1.2 Focus of Work

Different ambiguities clearly require different kinds of contextual information to be resolved. The senses that are linked to specific syntactic patterns are typically easier for people to distinguish. When sense distinctions are linked to the semantics of the verb's arguments, sense separation is often not so straightforward. In the present work, our goal is to model some of the processes through which semantics of arguments contributes to disambiguation. In particular, we would like to separate out and evaluate the contribution of a single argument position to sense differentiation for the verb, while eliminating the influence of other context elements.

Since the factors affecting the resolution of different ambiguities are often interdependent, automatic sense detection and induction systems are not usually designed to treat different kinds of sense distinctions separately. A number of sense-tagged corpora have been developed for the training and testing of such systems. This kind of annotation typically involves tagging each occurrence of a given word in text with a sense from a particular sense inventory. Sense inventories are usually taken out of machine-readable dictionaries or lexical databases, such as WordNet (Fellbaum, 1998), Roget's thesaurus (Roget, 1962), Longman Dictionary of Contemporary English (LDOCE, 1978), Hector dictionary, etc. In some cases inventories are (partially or fully) constructed or adapted from an existing resource in pre-annotation stage, as in PropBank (Palmer et al., 2005) or OntoNotes (Hovy et al., 2006). The quality of the annotated corpora depends directly on the selected sense inventory, so for example, SemCor (Landes et al., 1998) which uses WordNet synsets, inherits all the associated problems, including using the senses that are too fine-grained and in many cases poorly distinguished.

Such annotation is very labor-intensive and typically provides no way to distinguish between different kinds of sense distinctions. Nor does it usually address the question of what factors allow the speakers to identify a particular sense. Since the contribution of different context elements to the activation of each sense is unspecified, it becomes impossible to perform adequate error analysis for the automatic systems for word sense disambiguation (WSD) and word sense induction (WSI). That is, in both cases, it becomes difficult to track the types of sense distinctions detected more successfully by a given system.

This problem seems to be solved to some extent in the kind of context-sensitive annotation provided in the FrameNet corpus (Ruppenhofer et al., 2006) and in the CPA patterns. Both resources do endeavor to specify the context parameters relevant for sense distinction, but both are not sufficiently complete. FrameNet, which

proceeds with sense analysis frame by frame, often specifies only one out of several senses for each lexical item.[5] The CPA approach, which relies on full context analysis for each word, is painstakingly slow and consequently lacks coverage. Also, as we will see below, the requisite semantic information is very context-dependent and difficult to capture during lexicographic analysis.

In this work, we present an unsupervised learning algorithm for simultaneous clustering of the words selectionally similar to a given sense of the target verb and the arguments activating that sense. We also develop an exploratory data set that targets sense distinctions linked to semantics of a single argument, and analyze the issues involved in identifying such sense distinctions, both manually and automatically.

1.3 Approach

1.3.1 Clustering Algorithm

The idea that semantic similarity between words must be reflected in the similarity of their habitual contexts of occurrence is fairly obvious and has been formulated in many guises (including the "distributional hypothesis" (Harris, 1985), the "strong contextual hypothesis" (Miller and Charles, 1991), and even the much-quoted remark from Firth, on knowing the word by the company it keeps (Firth, 1957)). When applied to the case of lexical ambiguity, it leads one to expect that an ambiguous word will be used in the same sense in similar contexts. However, one of the main problems with applying the idea of distributional similarity in computational tasks is that in order to use any kind of generalization based on distributional information, one must be able to identify the sense in which a polysemous word is used in each case.

In CPA, the semantics of the arguments that help to differentiate a particular verb sense is often represented by the lexicographer as a *lexical set*, i.e. a collection of lexical items unified by a particular semantic feature (Hanks, 1996; Rumshisky et al., 2006). We follow this idea in developing an automated method for clustering the arguments of a predicate according to the sense they activate. Since our main focus is on lexical mechanisms at work, we restrict ourselves to modeling the contribution of NP heads, rather than full noun phrases, to disambiguation (i.e. we only use binary dependencies). We later discuss how the same method can be used in identifying semantic contribution of full NPs (cf. Ch. 7).

The clustering method we propose uses the notion of *contextualized similarity*.

[5]For instance, out of 20 fairly frequent verbs we surveyed (cf. Appendix C), only 7 had their main sense distinctions captured in FrameNet. Only 25 out of the total 70 identified senses for these verbs had a corresponding link to a FrameNet frame. Common verbs such as *assume, claim, cut, deny, enjoy*, and *launch*, had only one out of two or three main senses represented in FrameNet.

Whereas two lexical items may not be distributionally similar overall, in a particular context they may be essentially equivalent. This equivalence is in terms of the aspect of meaning they select. This applies to both sides of the predicate-argument relation. Selection is a *bidirectional* process: a given noun in the specified argument position activates a particular interpretation for the target verb, while a given interpretation of the target verb selects for a specific semantic component in an argument (cf. Ch. 3).

For example, the verbs *rub* and *dip* denote very different actions, but they both select for physical objects in direct object position. The same is true of *tackle* and *handle* which, in one of their senses, both select for the [+problem] component in the direct object position. At the same time, their arguments, for example, *emergency* and *job*, while not synonymous or semantically related, are similar in that they carry (or are assigned through coercion) the required semantic component [+problem]. This is also the case for the words that co-occur with polysemous nouns such as *lunch*. For example, *cancel* and *attend* each have a very different set of senses, and their frequencies of occurrence do not have a similar distribution across contexts. However, with respect to taking *lunch* as direct object, they are quite similar: they both select for the EVENT, rather than FOOD interpretation.

We use the notion of *selectional equivalence* to capture this intuition. Our clustering method (cf. Ch. 4) relies on contextualizing the representation of each lexical item to a particular target context. Selectional equivalents for each sense of the target verb induce clusters of nouns activating that sense.

1.3.2 Sense-Annotated Data Set

In the past few years, a number of initiatives have been undertaken to create a standardized framework for the testing of WSD and WSI systems, including the recent series of SENSEVAL competitions (Agirre et al., 2007; Mihalcea and Edmonds, 2004; Preiss and Yarowsky, 2001), and the shared semantic role labeling tasks at the CoNLL conference (Carreras and Marquez, 2005; Carreras and Marquez, 2004). Despite these efforts, there are still effectively no established criteria for the evaluation of the automated systems that deal with word sense detection. As mentioned above, the standard sense-tagged data sets conflate different kinds of contextual information. Different types of sense distinctions are also not treated separately. As a result, such data are not particularly well-suited for testing the discriminatory powers of an automated system. Evaluation schemes that compute the overall accuracy of sense detection systems over such corpora do not reflect the actual effectiveness of these systems. This situation is exacerbated by the fact that sense annotation has often been done on corpora which are not well-balanced, such as the Wall Street Journal data. As a result, the distribution of annotated instances does not reflect the

actual frequency distribution between senses, as evidenced by some of the data sets produced for the last several Senseval competitions. Especially for verbs, the most frequent sense often dominates the data set.[6]

The data set we develop is an attempt at addressing the problem of creating annotation targeting a specific factor contributing to predicate disambiguation. In this case, we target semantics of an argument in a particular argument position. The purpose of creating such a data set is two-fold. The first objective is to examine the way speakers deal with verbal ambiguities that depend on the semantics of the arguments. The second objective is to evaluate how well the clustering algorithm can perform inducing such senses.

The set of verbs we have chosen for annotation, along with the relevant argument position for each verb, was selected so that the sense distinctions in each verb's sense inventory could be detected by looking at semantics of the noun in the specified argument position. In several sense inventories, some of the included senses were clearly additionally influenced by other arguments, or to a large extent determined by the syntactic constructions. However, each sense inventory contains at least one sense pair that can be distinguished through the semantics of the noun in the specified argument position.

As will be discussed below, annotation of polysemous verbs in general, and especially sense distinctions dependent on semantics of the arguments, is plagued by the problem of *boundary cases*, where no clear sense assignment can be established, despite the fact that the relevant senses are well separated in most contexts. Consequently, we propose that only the annotation of clear cases be kept in the annotated data sets.

1.4 Practical Applications

Automatically inducing semantic preferences of polysemous verbs has a number of uses. The output of the algorithm we present can be used both for automatic sense detection and to assist in human analysis of selection phenomena. While our system does not target general purpose sense induction, one of its intended applications is to be used within a complete WSD or WSI system. The clustering solutions produced by our system can also be used for lexicographic purposes, both to examine selectional behavior of polysemous verbs and to enhance lexicographic tools that facilitate the task of sense definition, such as the Sketch Engine (Kilgarriff et al., 2004).

Specifying semantic requirements imposed on each other by the words entering a dependency relation can clearly be helpful in a variety of parsing tasks. For example, Gamallo et al. (2005) use a technique for clustering contexts with similar selectional

[6] see Ch. 6 for further discussion.

requirements to improve prepositional phrase attachment in Portuguese. Clusters of semantically similar words or manually constructed lexical hierarchies have also been used with success to improve both PP-attachment and NP-chunking, as well as semantic interpretation of noun compounds in English (Pantel and Lin, 2000; Rosario and Hearst, 2001). Using contextualized clustering should prove beneficial for such tasks.

In Chapter 7, we discuss some other applications of the presented algorithm, including the resolution of regular polysemy in nouns, as well as disambiguation of the noun phrases with semantically light head noun, based on the clustering of modifiers.

1.5 Outline

In Chapter 2, we discuss how the problem of sense inventory definition is treated in lexicographic and theoretical literature. We then review the use of different distributional similarity measures in different computational solutions to the task of establishing semantic similarity between words. In Chapter 3, we discuss the difficulties that arise in designing sense inventories for cases where predicate sense distinctions are strongly linked to the semantics of the arguments. We examine the issues that have to be addressed by automatic algorithms aiming to detect such sense distinctions. In Chapter 4, we present the *bipartite contextualized clustering* algorithm for clustering selectional equivalents of different senses of a polysemous verb and the corresponding semantically diverse arguments that activate each sense. In Chapter 5, we describe the development of an experimental data set that targets the semantics of a single argument as the deciding factor in detecting the correct verb sense. We discuss the impact of semantic relations between senses on the sense assignment decisions made by the annotators and propose some modifications to the standard sense annotation practices. In Chapter 6, we present an evaluation of the algorithm described in Chapter 4 in a standard word sense induction setting, using the data set described in Chapter 5. In Chapter 7, we discuss the applications of bipartite contextualized clustering to other cases requiring the resolution of polysemy.

Published Work

This thesis contains in part material from the following papers published or submitted for publication: (1) Rumshisky, A. and Grinberg, V. A. (2008). Using semantics of the arguments for predicate sense induction. (2) Rumshisky, A. and Batiukova, O. (2008). Polysemy in verbs: systematic relations between senses and their effect on annotation. In *COLING Workshop on Human Judgement in Computational Linguistics (HJCL-2008)*, Manchester, England. (3) Rumshisky, A. (2008). Resolving polysemy

in verbs: Contextualized distributional approach to argument semantics. *Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica*. (forthcoming) (4) Pustejovsky, J. and Rumshisky, A. (2008). Between chaos and structure: Interpreting lexical data through a theoretical lens. *Special Issue of International Journal of Lexicography in Memory of John Sinclair*. (forthcoming) (5) Rumshisky, A., Grinberg, V. A., and Pustejovsky, J. (2007). Detecting Selectional Behavior of Complex Types in Text. In Bouillon, P., Danlos, L., and Kanzaki, K., editors, *Fourth International Workshop on Generative Approaches to the Lexicon*, Paris, France.

Chapter 2

Related Work

2.1 Defining a Sense Inventory

Given that our goal is to model sense differentiation, it is important to understand what principles are used to create the inventory of senses. Creating sense inventories is a task that is notoriously difficult to formalize. This is especially true for polysemous verbs with their constellations of related meanings. In lexicography, "lumping and splitting" senses during dictionary construction – i.e. deciding when to describe a set of usages as a separate sense – is a well-known problem (Hanks and Pustejovsky, 2005; Kilgarriff, 1997; Apresjan, 1973). It is often resolved on an ad-hoc basis, resulting in numerous cases of "overlapping senses", i.e. instances when the same occurrence may fall under more than one sense category simultaneously.

This problem has also been the subject of extensive study in lexical semantics, addressing questions such as when the context selects a distinct sense and when it merely modulates the meaning, what is the regular relationship between related senses, and what compositional processes are involved in sense selection (Pustejovsky, 1995; Cruse, 1995; Apresjan, 1973). A number of syntactic and semantic tests are traditionally applied for sense identification, such as examining synonym series, compatible syntactic environments, coordination tests such as *cross-understanding* or *zeugma* test (Cruse, 2000). None of these tests are conclusive and normally a combination of factors is used.

These considerations have also been the concern of the computational community working on sense disambiguation, where evaluation requires having a uniform sense inventory. At the recent Senseval competitions (Mihalcea et al., 2004; Snyder and Palmer, 2004; Preiss and Yarowsky, 2001), the choice of a sense inventory frequently presented problems, spurring the efforts to create coarser-grained sense inventories (Navigli, 2006; Hovy et al., 2006; Palmer et al., 2007). Inventories derived from WordNet by using small-scale corpus analysis and by automatic mapping to top

entries in Oxford Dictionary of English were used in the most recent workshop on semantic evaluation, Semeval-2007 (Agirre et al., 2007). One of the proposed views has been that that it is impossible to establish a standard inventory of senses independent of the task for which they are used (cf. Agirre and Edmonds, 2006; Kilgarriff, 1997).

2.1.1 Lexicographic work

Establishing a set of senses available to a particular lexical item and (to some extent) specifying which context elements typically activate each sense forms the basis of any lexicographic endeavor. Several current resource-oriented projects undertake to formalize this procedure, utilizing different context specifications. FrameNet (Ruppenhofer et al., 2006) attempts to organize lexical information in terms of script-like semantic frames, with semantic and syntactic combinatorial possibilities specified for each frame-evoking lexical unit (word/sense pairing). FrameNet uses Fillmore's case roles to represent semantics of the arguments. Case roles (*frame elements*) are derived on ad-hoc basis for each frame. Context specification for each lexical unit contains such case roles (e.g. Avenger, Punishment, Offender, Injury, etc. for the Revenge frame) and their syntactic realizations, including grammatical function (Object, Dependent, External Argument (= Subject)), etc.), and phrase type (e.g. NP, PP, PPto, VPfin, VPing, VPto, etc.). Core frame elements represent semantic requirements of the target lexical unit, some of which may not be actually expressed in the sentence.

Semantically tagged data produced within computational community has often used available machine-readable dictionaries (MRDs) and lexical databases. For example, the SemCor corpus developed within the framework of Senseval competitions uses WordNet senses (Fellbaum, 1998) to tag a 700K word subset of the Brown corpus (Landes et al., 1998). An early Senseval competition used a set of senses from the Hector project for semantic tagging (Preiss and Yarowsky, 2001).

PropBank (Palmer et al., 2005) specifies verb senses in terms of framesets where each frameset consists of a set of semantic roles for the arguments of a particular sense of the target verb. A set of semantic arguments, numbered beginning with 0, is specified for each verb. Semantic roles are defined on a verb-by-verb basis, with the exception of the standard Agent and Patient/Theme assigned to Arg0 and Arg1, respectively. In the OntoNotes project, annotators use small-scale corpus analysis to create sense inventories derived by grouping together WordNet senses. The procedure is restricted to maintain 90% inter-annotator agreement (Hovy et al., 2006).

In the efforts to automate sense detection, the semantics of the arguments is frequently represented with information derived from external knowledge bases (e.g. WordNet, Roget, LDOCE, SUMO, various upper-level ontologies). However, very often, semantic components that activate the verb's sense require much more refined semantic grouping. In the annotation efforts that use semantic role labels to represent

semantics of predicate arguments (as in FrameNet, PropBank), such labels are usually added on as-needed basis, and systematizing the resulting set of semantic roles is very difficult.

This arbitrariness seems to be a common practice, motivated by variability of semantic requirements. For example, Church and Hanks (1990) discuss semantic tags that in effect assign contextual semantic interpretation to arguments of the verb *save*. They mention an eclectic set of semantic markers, including BAD, ENVIRONMENT, ANIMAL, MONEY, DESTRUCTION, ECON(OMIC), POLITICAL, INSTITUTION, LOCATION, PERSON, CORPORATION, MONEY, NUMBER. Church and Hanks (1990) remark that these tags were being extrapolated on the fly by a human from a set of concordances, and they in many cases correspond to the words strongly associated with *save* in the corpus. Pustejovsky et al. (2004) propose to use such lexical items to create promoted types in an ontology.

In CPA (Pustejovsky et al., 2004), such contextual semantic interpretations were systematized resulting in a set of shallow types, roles, and polarities, also derived on as-needed basis. This set is complemented by *lexical sets* when necessary. We review the main principles of Corpus Pattern Analysis below.

Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) (Hanks and Pustejovsky, 2005) attempts to catalog norms of usage for individual words, specifying them in terms of context patterns. Each pattern gives a combination of surface textual clues and argument specifications, offering a contrastive analysis of senses for each word. CPA uses the extended notion of syntactic frame which is understood broadly to include the following elements:

- argument structure
- minor categories: adverbial phrases, locatives, temporal adjuncts, purpose clauses, rationale clauses, etc.
- subphrasal cues: genitives, partitives, bare plural/determiner distinctions, infinitivals, negatives etc.
- collocational cues from wider context.

The semantics of the arguments is represented either through a set of shallow semantic types representing basic semantic features (e.g. Person, Location, PhysObj, Abstract, Event, etc.) or extensionally through *lexical sets*, which are effectively collections of lexical items.

As a corpus analysis technique, CPA derives from the analysis of large corpora for lexicographic purposes, of the kind that was used for compiling the Cobuild dictionary (Sinclair and Hanks, 1987). For each target word, a lexicographer sorts its contexts of usage into groups and records a pattern that captures the relevant semantic and

syntactic features, with a corresponding specification given for each group.[1] Several context patterns may represent a single sense, with patterns varying in syntactic structure and/or the encoding of semantic roles relative to the described event. A distribution of frequencies is associated with each sense and is typically very uneven. For example, CPA patterns for the verb *fire*[2] are given below.

Selected CPA Patterns for *fire*:

```
I DISCHARGE A GUN AT A TARGET (60%)

1.  [[Person]] fire [[Artifact=Firearm]] (at [[PhysObj]])

2.  [[Person]] fire [[Artifact=Projectile]] (off) (from [[Artifact=Firearm]]) (at [[PhysObj]] |
[ADV[Direction]])

3.  [[Person]] fire [NO OBJ] (at [[PhysObj]] | on [[HumanGroup]] | [ADV[Direction]])

4.  [[Artifact=Firearm]] fire [NO OBJ] (at [[PhysObj]] | on [[HumanGroup]] | [Adv[Direction]])

II DISMISS AN EMPLOYEE (11%)

5.  [[Person 1]] fire [[Person 2]] (for [[Action=Bad]])

III INSPIRE SOMEONE (11%)

6.  [[TopType]] fire [[Person]]'s [[Attitude=Enthusiasm]]

7.  [[TopType]] fire [[Person]] (up)
```

Many patterns have alternations, recorded in satellite CPA patterns. Alternations are different realizations of the same norm, rather than creative variations of that norm. However, alternations are linked to the main CPA pattern through the same sense-modifying mechanisms as those that allow for exploitations (coercions) of the norms of usage to be understood. For example, consider the set of patterns for the verb *treat*:

Selected CPA Patterns for *treat*:

```
1.  [[Person 1]] treat [[Person 2]] (at | in [[Hospital]]) (for [[Injury] | [Ailment]]); NO
[Adv[Manner]]

2.  [[Person 1]] treat [[Person 2]] [Adv[Manner]]

3.  [[Person]] treat [[TopType 1]] as | like [[TopType 2]]

4.  [[Person]] treat [[TopType]] as if | as though | like [CLAUSE]
```

[1] This process is referred to *triangulation* in Church and Hanks (1990) where in order to establish the meaning of a word, a lexicographer uses a collocate of that word, or another context element, such as a time adverbial, when studying concordances.

[2] Only patterns for the senses that are frequent enough to account for more than 5% of use are given.

Alternations for the first pattern are given below:

Alternations for Pattern 1 of *treat*:

```
[[Person 1]] treat [[Person 2]] (at | in [[Hospital]]) (for [[Injury | Ailment]]); NO [Adv[Manner]]

Alternation 1:  [[Person 1 <--> Medicament | Med-Procedure | Institution]]

Alternation 2:  [[Person 2 <--> Injury | Ailment | Bodypart]]
```

Table 2.1 gives the BNF specification for a CPA pattern grammar. Round brackets indicate optional elements of the pattern, and curly brackets indicate syntactic constituents. This specification relies on word order to specify argument position, and is easily translated to a template with slots allocated for each argument. Within this grammar, semantic roles can be specified for each argument.

2.1.2 Theoretical work

In lexical semantics, a number of attempts have been made to define the necessary structures for a lexical entry in order to account for regular polysemy, as well as model the processes involved in sense modification (Pustejovsky, 1995; Cruse, 2000; Apresjan, 1973; Mel'chuk, 1982).

Cruse (2000) summarizes some of the common ideas related to these problems. He views *antagonism* between readings as a defining criterion for ambiguity of linguistic expression. This applies equally to polysemy and homonymy. Resolution of ambiguity in context involves *selection* of one of the word's senses by the context. If the particular sense required by the context is not available from the lexical item, *coercion* occurs, and the required reading is created through *sense extension* processes such as metaphor or metonymy. In this view, between monosemy (a single sense) and polysemy (multiple senses), there is a number of intermediate cases which do not qualify for 'full sensehood'. These include 'facets' and 'perspectives',[3] which are non-antagonistic; and also 'subsenses'. All of these cases display different degrees of discreteness and are distinct from contextual modulation of the same sense. In sense modulation, a particular aspect of meaning is highlighted, while other aspects are suppressed or obscured.

A similar view of the nature of ambiguity was taken by Apresjan and others from the Moscow School of Semantics in their approach to 'explication' of meaning within an explanatory dictionary (cf. Zholkovsky et al., 1961; Mel'chuk, 1974; Apresjan, 1973; Apresjan, 1974; Mel'chuk and Zholkovsky, 1984; Apresjan, 2000). In this view, separation between meanings of a word is also seen as a matter of degree: from com-

[3]These phenomena are modeled within Generative Lexicon as *complex types* and *qualia roles*, respectively. We discuss them in more detail below.

pletely discrete in case of homonymy, to somewhat related in case of metaphorically motivated polysemy or other non-immediate polysemy, to more related in case of functional polysemy, to the immediate polysemy (e.g. Rus. *vyparit' sol'* "boil out the salt (from water)" vs. *vyparit' pyatno* "boil out the stain (as from fabric)") where the separation of meanings can very easily be contested. Lexical ambiguity is viewed as being on a scale, with homonymy on one end and monosemy on the other. Monosemy is a range on this scale, rather than a point: certain cases of monosemy may approach polysemy (e.g. Rus. *gasnut'* "cease to burn or shine" – there is an 'inclusively disjunctive organization of semantic components', but a coordination test suggests that it is the same meaning: *drova v kamine i fonari na ulice pogasli pochti odnovremenno* "the firewood and the street lamps went out almost simultaneously").[4]

According to this approach, a 'lexicographic portrait' of a word must include information about the interaction of different facets of a lexeme, as well as information about its government and co-occurrence properties. Combinatorial properties of a given word, or 'lexical co-occurrence constraints', are specified in the form of a list of words which form syntactic dependencies with the word in question and which share a particular semantic property. A dictionary entry contains information about linguistic features, exact and inexact synonyms, hypernyms, and derivatives. Lexemes are organized as 'lexicographic types', which are groups of lexemes with shared properties that are accessed and used by some grammatical or other general linguistic rules (e.g. "length, height, width, thickness, depth", cf. Apresjan (2000), p. 236) A dictionary entry provides a description of the ways in which a lexeme differs from other members of its lexicographic type. Lexicographic types are not disjoint, but rather they are seen as repeatedly intersecting. The same lexeme may appear in different classes associated with any of its properties. Semantic associations in language which allow metaphorization ('associative features') must also be included in the dictionary (e.g., for *lightning*, that it "may be associated with quickness and brilliance", etc.).

In this approach, lexical meanings are decomposed into simpler semantic components. Complex meanings are 'gradually reduced' to language-specific semantic primitives. The word's meanings are described in a special metalanguage whose vocabulary consists of semantic primitives and 'intermediate concepts' that can be

[4]This is similar to the *zeugma* and *cross-understanding* tests, the coordination tests discussed in Zwicky and Sadock (1975) and Lascarides et al. (1996). These tests check whether two coordinated conjuncts may activate different meaning of the ambiguous word. If such constructions seem acceptable to the speakers and do not create a feeling of a joke or a pun, then the word is judged to have a single underspecified meaning. Otherwise, two senses are postulated. For example, "*Teachers* are allowed to take maternity or paternity leave" sounds neutral, but "John *expired* the same day as his driver's license" evokes the perception of word play. Consequently, *teacher* is seen as underspecified for gender, but *expire* is seen to have two senses. The difference between cross-understanding and zeugma is that in the latter case the conjuncts have a strong preference for one of the interpretations (as is the case with *maternity* and *paternity*).

reduced to primitives in one or more steps. Each 'explication' of meaning has a hierarchical organization, i.e. each meaning is described by means of at least two semantic blocks, so a gradual breakdown is achieved of the more complex senses into the more simple ones (cf. Apresjan, 2000, p. 219).[5]

Lexical polysemy is viewed as a 'capacity of a word to have related meanings', i.e. meanings that have 'non-trivial common components' either in their definition tree (aka 'semantic tree') or in their 'associative features'. A word is defined as polysemous if for any two meanings of that word there exists a chain of related meanings that links them. Polysemy can therefore be radial or concatenated. Polysemy of a given word is regular if it has related meanings such that there is at least one other word that has the meanings that are related in exactly the same way. In metonymically or metaphorically motivated polysemy, metaphorization is achieved through suppressing or replacing one of the components of meaning. Metaphoric transfers tend to create irregular polysemy. Any non-immediate polysemy is also usually irregular.[6] Metonymic transfers, such as figure/ground, container/containee, organization/location, process/result, etc. tend to create regular polysemy. Other processes that tend to create regular polysemy include 'semantic analogy', 'compression of phrases' (i.e., elliptical omission is one of the mechanism of sense formation, e.g. Rus. *mashinka (dlja brit'ja)* "razor" vs. *(pishushchaja) mashinka* "typewriter"), and 'word-formation processes'. One of the criteria for the existence of different meanings is the existence of different derivatives of the same word that are morphologically similar, but have distinct meanings.

The notion of 'lexical function' is introduced to account for the systematicity of meaning transformations within language and to capture the regularity in lexical co-occurrence and derivation phenomena (Mel'chuk, 1974; Apresjan et al., 1969; Mel'chuk, 1982; Mel'chuk, 1996). A lexical function expresses a relation that holds between a pair of lexical items, i.e. lexical functions of the form $f(X) = Y$ denote a lexical-semantic relation that holds between the headword X and the value Y (its collocate). Lexical functions capture both syntagmatic and paradigmatic relations,[7] as illustrated in (2.1)-(2.4).

(2.1) **Mult**(flowers) = bunch
Mult(sheep) = flock
Mult(dog) = pack

[5]This view is similar to that of Wierzbicka who also implements the notion of semantic primitives, with the main difference being that she considers them universal, rather than language-specific, and not necessarily hierarchically organized.

[6]Non-metaphoric immediate polysemy can also be irregular, for example, Rus. *podnozhka* is ambiguous between "footstep" and "a blow on the legs that trips the opponent", where *noga* ("foot/leg") is the common component.

[7]These examples are taken from Apresjan et al., 1969; Mel'chuk, 1982; Fontenelle, 1998.

(2.2) **Magn**(bachelor) = confirmed
Magn(pain) = excruciating
Magn(fear) = mortal
Magn(contrast) = sharp, vivid

(2.3) **Anti**(beautiful) = ugly
Anti(friend) = foe, enemy
Anti(before) = after
Anti(love) = hate
Anti(to open) = to close

(2.4) **Gener**(fluid) = substance
Gener(blue) = color
Gener(crawl) = move

Standard lexical functions express meanings that are very general and which can be lexically expressed in a variety of ways.

Generative Lexicon

Generative Lexicon (GL) (Pustejovsky, 1995; Pustejovsky, 2001; Pustejovsky, 2006; Pustejovsky, 2008) introduces a number of mechanisms for modeling compositional behavior of words and the accompanying meaning transformations. Each lexical entry is a complex data structure that contains the information about event and argument structure, as well as qualia roles. The former play a larger role in the lexical semantic interpretation of events, while the latter are more important for the modeling of the compositional behavior of nouns. Qualia structures within a lexical entry allow for a number of generative mechanisms to produce sense modification. In the classic GL model, there are four qualia roles:

AGENTIVE quale specifies the origin of the object
TELIC quale specifies its purpose or function
CONSTITUTIVE quale specifies its constituent parts or material
FORMAL quale specifies its place within a larger domain

SIMPLE (Busa et al., 2001; Lenci et al., 2000), an ontology based on GL principles, provided an extended set of qualia which included subtyping for each of the main qualia types, including DIRECT, INDIRECT, or INSTRUMENT TELIC, as well as DIRECT or INDIRECT AGENTIVE, and some others. Thus, different types of TELIC quale distinguish whether the entity is the object of its intended activity (as *drink* is for

beer), the subject (as *play the drums* is for *drummer*), or the instrument of that activity (as *cut* is for *knife*).

The qualia play an important role in modeling the processes through which argument typing constraints of the predicates are satisfied in composition. Qualia bindings for a particular noun license its occurrence in coercive, type-shifting contexts where the type expected by the predicate is satisfied by the noun through the use of the qualia role. For example, a predicate that requires an event in the direct object position, such as *begin*, may coerce an artifactual entity, such as *sandwich* to one of the event values in its qualia structure, in this case, the TELIC *eat* (since *sandwich*'s intended function is to be *eaten*).

The meaning of adjectival modifiers is also determined with respect to the qualia role that a particular adjective modifies. Thus, for example, *a good knife* is a knife that cuts well, while *a good meal* is a meal that tastes well. The meaning varies, but in both cases the interpretation for the adjective is generated because it acts on the telic quale of the noun.

The qualia structure of a noun varies depending on whether it is a *natural*, a *functional or artifactual*, or a *complex type* (Pustejovsky, 2001). The natural types are the naturally occurring objects and phenomena, such as *rock*, *sun*, *light*, etc. The functional, or artifactual, types are the artifactually constructed objects (such as *beer* or *knife*) that typically have an intrinsic purpose or function. This distinction between the type levels is made for all major categories in the language, so for example, there are natural predicates such as *fall* and artifactual predicates such as *fix*. Complex type is a term used for concepts that combine two (or more) distinct semantic types, each with its own set of qualia roles. Complex types are introduced in GL as a mechanism for dealing with selectional behavior of nouns such as *lunch* (EVENT • FOOD) and *newspaper* ((PHYS • INFO) • ORGANIZATION). In these cases, a separate

qualia structure is associated with each of the dotted types.[8]

Several types of compositional processes are distinguished in argument selection. Broadly, three categories are possible:

1. *Pure selection* (*type matching*), or *accommodation* (*subtype coercion*). This mechanism operates when the type expected by the predicate either matches directly or is inherited by the type of the argument.

   (2.5) a. The rock fell to the floor. (*pure selection*, PHYSICAL OBJECT)
   b. Mary drove a Honda to work. (*accommodation*, HONDA $\sqsubseteq$ CAR)

   (2.6) a. The food spoiled. (*pure selection*, PHYSICAL OBJECT with TELIC *eat*)
   b. John read the book. (*pure selection*, PHYSICAL OBJECT • INFO)

2. *Coercion.* This group of mechanisms operates when there is a type mismatch, i.e. the type of the argument does not match the type expected by the predicate. *Type coercion* can be *domain-preserving* (e.g. an entity is coerced into another entity type) or *domain-shifting* (e.g. an entity is coerced into an event type). Two kinds of selection mechanisms are distinguished in *type coercion*:

   - *Exploitation.* In this operation, a subcomponent of the argument's type is accessed and exploited. For example, a component type of the complex type or the base type of an artifactual might be selected:

     (2.7) a. Mary threw the knife.

[8] Component types of the complex type can be seen as fully distinct but non-antagonistic readings of a word (e.g. *book* "text" vs. *book* "tome"). Cruse (2000) summarizes their properties as follows:

- They can occur simultaneously (e.g., "publish a book");
- They can be metaphorically extended simultaneously (e.g., "your mind is an open book");
- Coordination tests fail to produce a sense of punning (e.g. "the book was badly written, but beautifully printed";
- They behave to a large extent as independent senses, which includes having:
  - independent truth conditions, (e.g. "did you like the book?" "yes, it is interesting"/"no, the print quality is horrible"),
  - independent metaphorical extensions (e.g., "book of matches" refers to the tome, and not to the text),
  - independent proper names (e.g. "David Copperfield" refers to the text and not to the tome),
  - independent components/meronyms and hyponyms (*novel, bibliography, dictionary* are seen as hyponyms of text, *paperback, hardback* as hyponyms of tome; *chapter, paragraph* are meronyms of text, *cover, page, spine* are meronyms of tome).

(accessing PHYSICAL OBJECT base type of an artifactual)
b. John believed the book.
(accessing INFORMATION component of a complex type)

- *Introduction.* This mechanism operates when the type required by the predicate is richer than the type of the argument. This operation then wraps the supplied type with the interpretation needed to satisfy the typing requirement of the predicate.

(2.8) a. The water spoiled. (a *natural type* is wrapped with a TELIC role and raised to an *artifactual*)
b. John read the wall. (an *artifactual type* is assigned an interpretation of a *complex type*)

Domain-shifting coercion operations can also access the available typing of the argument or wrap the argument's type to introduce the required interpretation:

(2.9) a. John enjoyed the beer. (*Exploitation*)
b. The authorities denied the attack. (*Introduction*)

A related set of non-coercive compositional mechanisms are involved in *Selective binding*, which accounts for the polysemy of adjectival modification. *Selective binding* also operates by accessing the qualia structure of the head noun, but it is non-coercive with respect to the noun type. Thus, adjectives that function as event predicates (such as *fast, long, good*) invoke selective interpretation of an event expression contained in the qualia for the head noun:

(2.10) a. a fast car (TELIC: to be driven)
b. a fast typist (TELIC: to type)

Similarly, the qualia structures of the head noun are accessed by the psychological predicates such as *happy, sad* that predicate over animate objects and time intervals, by the adjectives that predicate over individuals or locations, such as *noisy*, and so on.

3. *Co-composition.* This mechanism operates when the sense of an expression is constructed by virtue of all constituent elements behaving as functors, as in "bake a potato" vs. "bake a cake", where depending on the complement, the base sense of *bake* is interpreted as either *change of state* or *creation.* The latter

case occurs when the agentive quale of the complement involves the governing predicate, i.e. when the object denoted by the complement is typically created by *baking*.

These mechanisms for modeling compositional behavior of words provide some of the tools necessary for a more refined analysis of sense selection processes.

2.2 Distributional Models for Semantic Similarity

The idea that distributional similarity can be used to determine semantic similarity has been used in a number of research tasks in natural language processing. These include, most notably, areas such as word sense induction (WSI), automatic thesaurus construction, word sense disambiguation (WSD), selectional preference acquisition (SPA), and semantic role labeling (SRL). In WSD, distributional similarity is used to group together occurrences of the same word according to the sense in which the word is used. In WSI and thesaurus construction, distributional similarity is used to obtain clusters of similar words. The clusters can be hard or soft, with soft clusters typically assigning words to multiple clusters with different probabilities.

Resulting distributional clusters are seen as a means to address the problem of data sparsity faced by many NLP tasks. The problem is that a lot of fairly common content words occur very infrequently in actual texts. Their counts thus can not be used to reliably predict their behavior, which is especially problematic since a significant percentage of actual texts is made up of precisely such rare events. (Dunning, 1993) reports, for example, that words with frequency of less than one in 50,000 make up 20-30% of news-wire reports. With respect to word cooccurrence, the problem is exacerbated further, since the number of possible joint events is much larger than the number of events actually encountered in texts. Generalizing across clusters allows us to model rare events, thereby alleviating the problems caused by sparsity in "middle layer" NLP tasks, including, for example, any number of parsing-related problems, such as resolving PP-attachment, scope of modification, nominal compounds, etc.

One of the main challenges in using distributional similarity to generalize over word classes is that one needs to resolve the problem of polysemy with respect to distributional representations. The problem is that for a polysemous word, we want the generalizations to apply to its different *senses*, rather than to all of its occurrences uniformly. In the absence of a semantically tagged corpus, obtaining frequency counts for each sense in a straightforward manner is impossible.

For example, in trying to acquire selectional preferences for verbs, (Resnik, 1996) had to rely on normalization by the number of senses each noun had, without actually being able to tell what sense of the target noun a particular verb occurred with in each instance. (Resnik, 1996) conceptualized noun senses as classes in a manually

constructed conceptual taxonomy (WordNet). Selectional preference of a given verb toward a particular noun class was modeled as *selectional association (SA)*, a measure based on *relative entropy* $D(p||q)$ (Cover and Thomas, 1991) between the probability distribution on semantic classes conditioned on the governing verb, and the corpus-wide probability distribution on the same classes irrespective of the syntactic context:

$$SA(v_i, c) = \frac{P(c|v_i) \log \frac{P(c|v_i)}{P(c)}}{D(P(c|v_i)||P(c))} \tag{2.11}$$

where $v_i \in V$ denotes a particular verb, and $c \in C$ is a given semantic class of nouns.

In order to use this measure, one needs to have estimates of joint probabilities $P(v, c)$, which (Resnik, 1996) obtains by dividing the total number of times a given noun occurs with a particular verb by the number of senses that noun has, and adding these up for all the nouns in a semantic class:

$$P(v, c_i) = \frac{1}{N} \sum_{n \in c_i} \frac{\text{freq}(v, n)}{|\{c : n \in c\}|} \tag{2.12}$$

where N is the total number of verb-noun pairs (v, n) extracted from the corpus[9]. This normalization aims to account for the fact that a noun may have more than one sense, and we only want to consider the counts of the relevant sense. But it is clearly inadequate, as a frequency distribution on the senses of a polysemous word is not typically uniform, and a word typically isn't equally likely to be used in any of its senses.

The same problem of polysemy has to be resolved in other computational tasks that rely on distributional similarity, and it is addressed directly in word sense induction, automatic thesaurus construction, and WSD. Consequently, there is a number of solutions as to how to access this information about senses computationally. The approaches typically differ along several dimensions, including: (1) the target task they try to address, (2) the context representation they adopt, (3) similarity measures they use, and (4) their methods for identifying and grouping together similar senses.

2.2.1 Context representation

First, since the driving idea is that words with similar senses are found in similar contexts, one has to consider what counts as context. A word is represented by a set of contexts in which it occurs in a corpus. This representation is typically viewed as a feature vector, where each feature corresponds to some context element. The value of each feature is the frequency with which that element is encountered together with

[9]Resnik (1996) experimented with selectional preferences for object position, so these were verb-object pairs.

the target word. Solving the problem of polysemy amounts to separating out the occurrences corresponding to each sense from a distributional vector that represents the target word. An alternative way to view distributional representation of a word is to treat it as a probability distribution on joint events of occurrence of the target word with each context element. Also, vector-based representation has a set-theoretic variant where frequency counts are replaced with 1's and 0's, depending on whether the context element (feature) ever co-occurred with the target word. Distributional vectors are then reduced to sets, each set being a collection of features. Yet another way is to regard each word, including the target word, as nodes in a co-occurrence graph, where the co-occurring context elements are represented by the neighboring nodes, and the frequency of the co-occurrence is the weight assigned to the corresponding edge (Widdows and Dorow, 2002; Agirre et al., 2006; Agirre and Soroa, 2007; Véronis, 2004).

Approaches differ with respect to which elements of the context are considered relevant. Some approaches use distributional features based on bag-of-words style co-occurrence statistics (Schütze, 1998; Gale et al., 1993; Widdows and Dorow, 2002), where context is represented by features that track the frequencies with which other words and/or small n-grams occur within a small window of the target word. Local features typically use a smaller window, topical features may track keywords occurring within a sentence or a paragraph. Other approaches use context representations that incorporate syntactic information, and sometimes semantic information from external sources. Such approaches use frequency counts of grammatical relations (GR) in which the target word participates, where each distributional feature corresponds to a grammatical relation and to the collocate or collocates linked to the target word by this grammatical relation. For example, the target word could be a noun that occurs as direct object to the verb *become*, serves as an indirect object to the verb *grow* inside a prepositional phrase introduced by *into*, and in some contexts governs a prepositional phrase *with hair* (as in, "become an obsession", "grew into an obsession", "obsession with hair"). If we represent each feature as a tuple containing the grammatical relation and the collocate or collocates, the features that will appear in its distributional representation might look like this:

$$\langle (become, \text{obj}), (grow, \text{iobj}, into), (hair, \text{mod}, with), ... \rangle$$

The chosen context representation may include only a particular set of grammatical relations, for example, only those corresponding to noun modifiers and subject-object relations with verbs (Hindle, 1990; Pereira et al., 1993). Alternatively, context representation may cover a full set of syntactic relations (Grefenstette, 1994; Lin, 1998; Pantel and Lin, 2002; Kilgarriff et al., 2004; Curran, 2004; Gamallo et al., 2005). In both GR and and non-GR-based approaches to context representation, the features may also track some other information about the collocate, for example, its semantic

class as given by some external source, its POS category, and so on.

2.2.2 Similarity measures

The second aspect that distinguishes between different approaches to distributional similarity is how the similarity measure itself is defined. Typically, raw frequency counts for each feature are normalized in some way to account for overall frequency of the target word and the feature-defining collocate. This normalization, or "weighting" (Curran, 2004), sometimes also aims to account for how strongly the target word is associated with the collocate. For example, the *association score* between the target word and the context element may be defined as mutual information between the two. Similarity measure itself is then computed using such association scores, rather than raw frequency counts. Defining a similarity measure thus entails selecting (1) a normalization scheme (or an *association score* used in constructing a distributional representation of each word, and (2) a measure of similarity between such representations.

Different distributional representations lend themselves to different definitions for similarity measures. Thus, there are vector space similarity measures (e.g. *cosine*, *Euclidean distance*, L_1 *norm*, overlap-based measures that use set-theoretic representation (*Dice*, *Jaccard*, etc.), graph-based similarity measures (e.g. node *affinity score* (Widdows and Dorow, 2002)), and information-theoretic measures based on the probability distribution representation (*relative entropy*, *Jensen-Shannon divergence*, *α-skew divergence*). Dagan (2000) gives a good overview of different similarity measures. Different proposals for computing distributional similarity are also summarized in Manning and Schütze (1999), in Ch. 4.2 of Curran (2004), and described in Lin, 1998; Lee, 1999; Weeds et al., 2004; Weeds and Weir, 2005, and elsewhere.

Some of the above similarity measures are summarized in Table 2.2 below. We assume the following notation: given two words, w_1 and w_2, $\vec{X}$ and $\vec{Y}$, respectively, will denote the feature vectors representing w_1 and w_2. We will use x_i and y_i to denote the *association scores* of w_1 and w_2, respectively, for the individual ith feature. A and B will denote the set-theoretic representations of w_1 and w_2 , collections of context elements (or features) extracted with each word anywhere in the corpus. The set-theoretic representation corresponds to a vector representation where the vector dimensions with non-zero frequencies get the value of 1, with the other dimensions having the value of 0. Finally, we will use p and q to denote the probability distributions associated with context elements (features) found to occur together with w_1 and w_2, respectively, with p_i and q_i denoting the probability of the ith feature, and $1 \geq i \geq |A \cup B|$.

Jaccard and *Dice* measures compute set-theoretic overlap. *Dice* measure gives the percentage of common features with respect to the average size of the set of all context

features found for w_1 and w_2. *Jaccard* measures the same percentage with respect to the union of two context sets. Both these measures are easily generalized from their set-theoretic definitions for the case of real-valued normalized frequency counts, where intersection corresponds to min and union to max (cf., for example, generalization of *Jaccard* in Grefenstette (1994))[10]. The *cosine* measure (which corresponds to the *correlation coefficient* in terms of probability distribution representation when association scores are appropriately normalized), as well as *Jaccard* and *Dice* measures have a range between 0 and 1. The *relative entropy* $D(x||y)$ (*Kullback-Leibler*, or *KL divergence*) (Cover and Thomas, 1991, p. 18) is asymmetric, and *Jensen-Shannon*, or *JS divergence* generalizes *relative entropy* to define a symmetric measure of divergence for two distributions, measuring their average divergence to a mean. *JS divergence* bypasses the well-known problem with *relative entropy*, namely, that when $\exists i : q_i = 0$ and $p_i \neq 0$, it gets a value of ∞. *JS divergence* has a range between 0 and $2 \log 2$. Neither *JS divergence*, nor *relative entropy* satisfies the triangle inequality ($f(x, y) + f(y, z) \geq f(x, z)$). The *$\alpha$-skew divergence* is another, asymmetric generalization of the *relative entropy* (Lee, 1999). And finally, it's also worth noting that *Euclidean distance* (or L_2 *norm*) and L_1 *norm* (or *Manhattan norm*) belong to the same class of geometric distances L_∞, but the latter is more frequently interpreted in probabilistic terms as "the expected proportion of different events" between two distributions (cf. Manning and Schütze, 1999; Curran, 2004)

2.2.3 Proposals for sense detection

In the present work, we are mostly interested in the approaches that use features based on grammatical relations, since the true path to identifying senses seems to lie through knowing syntactic structure into which the target word's collocates are embedded. We will now briefly review a few of the relevant approaches in the literature, characterizing them with respect to the above aspects, that is, the context representation used, the adopted similarity measure(s), the target task addressed, and where applicable, the methods for grouping together similar senses.

Pointwise mutual information as the association score

Church and Hanks (1990) proposed to use mutual information[11] to identify words strongly associated with each other. This measure considers only direct association statistics between two words, and does not aim to account for their co-occurrence with other words. It has since been frequently used as the association score between context feature and the target word in similarity computations. Mutual information $I(w_1, w_2)$

[10] The generalizations are marked with † in Table 2.2

[11] A more contemporary term is *pointwise mutual information* (Cover and Thomas, 1991; Manning and Schütze, 1999).

is defined as the log likelihood ratio of the observed probability of co-occurrence of the two words, as compared with the probability of co-occurrence expected by chance (i.e., by assuming independence):

$$I(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \tag{2.13}$$

A well-known criticism of the mutual information measure is that in case of pronounced dependence, where $p(w_1) \approx p(w_1, w_2)$, the mutual information value becomes unreasonably high for infrequent events (i.e. for small values of $p(w_2)$):

$$I(w_1, w_2) = \log \frac{p(w_1)}{p(w_1)p(w_2)} = \log \frac{1}{p(w_2)} \tag{2.14}$$

Church and Hanks (1990) point out that mutual information can be used to determine candidacy for strong association under any desired conditions, including POS markup, syntactic relations, arbitrary entity identification ("such as 'person', 'place', 'time', 'body-part', 'bad', etc."), and so on.

Classes of semantically similar nouns based on their distribution as subjects and objects of verbs

Hindle (1990) produced classes of semantically similar nouns, using their grammatical relations with verbs to represent context. He restricted possible contexts for nouns to subject and object relations within a clause. Mutual information was used as the association score between the verb and the noun in the appropriate argument position. Using these association scores, Hindle (1990) defined separate *subject* and *object similarity* for two nouns with respect to a particular verb, summing the two similarity values over all verbs to obtain the distributional similarity measure.

Association scores were computed frequency counts from a parsed corpus:

$$A_{obj}(n, v) = \log \frac{p(v, n)}{p(v)p(n)} = \log \frac{freq(n, v)/N}{freq(n)/N * freq(v)/N} \tag{2.15}$$

where N denotes the total number of clauses extracted from the corpus. *Object* (and similarly, *subject*) *similarity* score of two nouns with respect to a particular verb was defined by looking at the whether each noun had a positive or negative association score, i.e. whether it occurred as the object of that verb more or less frequently than expected by chance. If both nouns associated with the verb "in the same direction", so to speak, the *object similarity* score was positive and equal to the "overlap" (i.e. the minimum) of the two scores. Otherwise, the *object similarity* was considered zero:

$$sim_{obj}(v, n_1, n_2) = \begin{cases} min(A_{obj}(v, n_1), A_{obj}(v, n_2)) & A_{obj}(v, n_1) > 0, A_{obj}(v, n_2)) > 0 \\ min(|A_{obj}(v, n_1)|, |A_{obj}(v, n_2)|) & A_{obj}(v, n_1) < 0, A_{obj}(v, n_2)) < 0 \\ 0 & otherwise \end{cases} \tag{2.16}$$

The similarity score is obtained by summing the obtained similarity values for the two argument positions over all verbs in the corpus:

$$sim(n_1, n_2) = \sum_{i=0}^{N} sim_{obj}(v_i, n_1, n_2) + sim_{subj}(v_i, n_1, n_2) \tag{2.17}$$

The aim of Hindle's study was to investigate the feasibility of using distributional information to obtain a useful semantic classification. He clusters nouns together by selecting, for each target noun, the top-10 most similar nouns. Though this method produces some semantically coherent sets of nouns, he remarks that further means of automatic discrimination would be necessary to filter out spurious clusters. He also looks at "reciprocally most similar" nouns to produce a list of likely synonyms (and sometimes, antonyms).

Probabilistic clustering of nouns into a pre-set number of sense clusters, based on their distributions as direct objects

Pereira et al. (1993) attempted to capture "hidden sense classes", that is, to obtain a direct representation for word senses using distributional information. They modeled senses that exist in a language as a set of "soft" word clusters with membership probability distribution assigned to each word. In particular, they tried to discover noun senses using the statistics of their co-occurrence with verbs in direct object position. Unlike Hindle (1990), their context representation does not track the verbs with which the nouns occurred in subject position, or any other context features. A noun n is represented by the *conditional distribution over verbs* with which it occurs in direct object position: $p(v|n) = \frac{freq(v,n)}{freq(n)}$, where $freq(n) = \sum_v freq(v, n)$. These distributions are clustered so as to produce a set of probabilistic clusters $\mathcal{C}$ with *cluster membership probabilities* $p(c|n)$ defined for each noun with respect to all clusters $c \in \mathcal{C}$. Each cluster's centroid is represented by a conditional distribution over verbs $p(v|c)$, obtained as a sum of the corresponding probabilities for each of its member nouns, weighted by each noun's contribution to the cluster[12]:

[12]Since clusters are probabilistic, the summation is over all nouns in the training set.

$$p(v|c) = \sum_{n \in \mathcal{N}} p(n|c) * p(v|n) \tag{2.18}$$

They use *KL divergence* $d(n, c) = D(p(v|n)||p(v|c))$ (cf. Table 2.2) as a measure of similarity between the distribution over verbs induced by a noun and the distribution corresponding to a cluster centroid. Their algorithm never measures similarity between the distributions induced by two nouns, thereby avoiding the problem of zero denominator in KL divergence (cf. p. 27)

The problem of clustering nouns is then reduced to finding *cluster membership distributions* $p(c|n)$ for each noun and *cluster centroid distributions* over verbs $p(v|c)$ for every cluster, using the set S of pairs (n_i, v_i) as training data. The resulting set of clusters must be such that for each noun n, conditional distribution $p(v|n)$ could be approximated as a sum of conditional probabilities $p(v|c)$ for all sense clusters the noun n belongs to, weighted appropriately by the noun's contribution to each cluster:

$$\hat{p}(v|n) = \sum_{c \in \mathcal{C}} p(c|n) * p(v|c) \tag{2.19}$$

First, they find cluster membership probabilities $p(c|n)$ that maximize overall *cluster membership entropy* under the condition of fixed *average cluster distortion*. Then they use maximum likelihood estimation to find the *cluster centroid distributions* $p(v|c)$. The *average cluster distortion* is given by:

$$\langle D \rangle = \sum_{n \in \mathcal{N}} \bar{D}(n) = \sum_{n \in \mathcal{N}} \sum_{c \in \mathcal{C}} p(c|n) d(n, c) \tag{2.20}$$

where $\bar{D}(n)$ is the average of distances $d(n, c)$ between noun n and cluster centroids of all clusters to which it belongs[13]. The *membership entropy* is defined as the sum of cluster membership entropies for individual nouns[14]:

$$H = \sum_{n \in \mathcal{N}} H(p(c|n)) = \sum_{n \in \mathcal{N}} \sum_{c \in \mathcal{C}} p(c|n) \log p(c|n) \tag{2.21}$$

Note that merely maximizing the entropy would give a distribution of equally probable senses for all nouns, which would be quite inaccurate. Maximizing the entropy while fixing average distortion gives an expression for cluster membership probabilities $p(c|n)$ dependent on parameter β.

Using (2.19) to obtain $\hat{p}(n, v) = \sum_{c \in \mathcal{C}} p(c)p(n|c)p(v|c)$, the likelihood of the data

[13]Effectively, $p(c|n)$ gives the relative frequency for noun n of the sense corresponding to the cluster c

[14]In the original paper, equation (6) contains a typo.

set S is expressed as $P(S) = \prod_{(n,v)\in S} \hat{p}(n,v)$, with the log likelihood of the model given by:

$$l(S) = \log \prod_{(n,v)\in S} \hat{p}(n,v) = \sum_{(n,v)\in S} \log \sum_{c\in\mathcal{C}} p(c)p(n|c)p(v|c) \tag{2.22}$$

Using the values obtained previously for *cluster membership distributions* $p(c|n)$, the expression in (2.18) for the *cluster centroid distributions* $p(v|c)$ is obtained. By gradually increasing parameter β from initially very low values, Pereira et al. (1993) successively split the data into a hierarchy of pre-selected number of probabilistic clusters. The obtained set of clusters minimizes, for each value of β, the expression $F = \langle D \rangle - H/\beta$, which corresponds to maximizing the entropy and minimizing the average distortion.

This clustering model was evaluated on two tasks. In the first task, the resulting clusters were evaluated by comparing the model estimates for the conditional distributions over verbs induced by nouns ($\hat{p}(v|n)$, cf. (2.19)), and the same distributions $p(v|n)$ estimated directly from the corpus[15]. The authors use relative entropy between the two distributions as a measure of comparison, averaging it over all nouns in the data set, and report the values for the number of clusters ranging between 0 and 400. In the second task, selected verb-object pairs were deleted from the training set, and the model estimates for the corresponding $\hat{p}(v|n)$ and $\hat{p}(v'|n)$ were used to predict which of the verbs v and v' is more likely to take n as direct object. The predictions were then compared with the ones obtained by using deleted pair counts.

The data set for both tasks consisted of verb-object pairs extracted for 1000 most frequent nouns in the 44 million word AP newswire corpus. The authors also reported clustering 64 direct objects of the verb "fire", and 1000 most frequent nouns in a 10-million word Grolier's Encyclopedia corpus. Note that the total number of clusters pre-set within their model effectively represents the number of all possible senses of nouns that could be encountered in the corpus.

Distributional clustering based on grammatical relations between nouns, verbs, and modifiers

Grefenstette (1994) measured distributional similarity using grammatical relations between nouns, verbs and adjectives. The relation set covered included adjectival and nominal noun modifiers (ADJ, NN), nouns within prepositional phrases attached to nouns (NNPREP); subjects and objects of verbs (SUBJ, DOBJ), and nouns within

[15]Model estimates were obtained from the training data and compared with the direct estimates for training, test, and new data sets. New data sets were comprised by the nouns not selected originally. Cluster membership distributions $p(c|n)$ for them were estimated using *KL divergence* between $p(v|n)$ and $p(v|c)$ for pre-computed clusters.

prepositional phrases attached to verbs (IOBJ). However, these relations were collapsed in similarity computations, effectively leaving four types of relations: *subj*, *obj*, *iobj*, *mod*, where *mod* collapsed ADJ, NN, and NNPREP relations. The information regarding what kind of preposition introduced an indirect object or a modifier was thus stripped away. Grammatical relations involving adverbs and numbers were also not used (cf. pp 40, 42, 44-45, 47). Distributional similarity computations were performed mostly for nouns, with some results reported for modifiers. The words were grouped together using *similarity lists*. Such a list, i.e. a list of most similar words, was compiled for each word in the corpus. Within each list, the words were further subdivided into groups according to the degree of similarity to the target word.

A weighted Jaccard measure generalized for real values (cf. Table 2.2) was used to measure the overlap between the attribute sets (= context-based features) of the two words being compared. The association score for each attribute was computed as a product of global and local weights. Global weighting accounts for how distinctive an attribute is (i.e. whether it occurs with many words in the corpus)[16]. Local weighting is applied to account for the actual frequency of the attribute for the target word. Grefenstette used adjusted log of frequency for local weighting, $weight_{loc} = 1 + \log freq(w, \text{feature}_i)$.

The resulting similarity values were evaluated against the following sources: (1) human judgements on words closely associated to adjectives, (2) pseudo-synonyms resulting from randomly splitting all occurrences of a given word into two sets; (3) overlap between dictionary definitions; (4) Roget's thesaurus entries. Grefenstette (1994) also compares the performance of window-based co-occurrences features to the ones using syntactic information to find that the latter significantly outperforms the former for the first 600 most frequent words in the corpus. For less frequent words, window-based approach, which tends to extract many more attributes, outperforms syntactic approach in precision. These results suggest that "frequently occurring events can be more finely analyzed than rarer ones" (pp. 94-99). Some other common issues that emerge include the insufficient specificity and/or accuracy of the parse, as well as the stability of the obtained similarity distributions (cf. pp. 60-61). Grefenstette (1994) considers the following applications for distributional similarity: (1) query expansion, (2) enrichment of WordNet thesaurus (identifying the appropriate hypernym sense to aid the template-based discovery of hypernym/hyponym pairs, cf. pp.114-125), (3) word meaning detection, and (4) automatic thesaurus construction.

In order to identify word senses, Grefenstette used an extension of Hindle's idea of looking at *reciprocally nearest neighbors*. Two words that make each other's top-10

[16]The global weighting scheme as described in Grefenstette (1994) doesn't seem to be consistent with the stated parameters, and Curran (2004) notes that the global weights actually used by Grefenstette were different and more consistent with reported results. However, there seems to be a typo or an omission in corresponding formula, eq. (3.9) in Curran (2004).

most similar list are seen to define a "semantic axis" along which each of the two words may be interpreted. Each of the words is used to define a semantic dimension of the other word. If there is overlap between the two words' top-10 lists, the other words may be "attached" to the axis. Only the words that are more frequent than word defining the sense for the other word get so added (cf. p. 126). The latter is done an attempt to make the sense definition more general, and it makes for an asymmetry in the definition of the corresponding senses of the two words. It's worth noting that the only case where two words will form such an axis is when their distributional profiles are very similar, which suggests that either their dominant senses are similar, or that they have a number of similar senses with similar distributions. The second case should, of course, be much less likely, unless it involves something like two near-synonyms with identical regular polysemy. It's also worth noting that the idea that distributionally close words form sense-defining clusters was used by other people to define senses in a more general situation. For example, Pantel and Lin (2002) defines tight clusters (*committees*) which represent senses in his clustering-by-committee algorithm (see below). A pair of *reciprocally near neighbors* is effectively a prototype for such "committee".

Thesaurus construction using a full set of grammatical relations

Lin (1998) defined a similarity measure using a full set of grammatical relations extracted by a parser. Each grammatical relation was represented as a "dependency triple" (w, R, w'), where R denotes the relation that holds between the words w and w' (e.g. (*become*, *obj*, *obsession*), (*obsession*, *obj-of*, *become*)). The mutual information-based *association score* $I(w, R, w')$ for the triple was defined as the log likelihood ratio of the probability of co-occurrence of w and w', given R, $p(w, w'|R)$, vs. the probability of their conditional co-occurrence that would be expected by chance $p(w|R) * p(w'|R)$:

$$I(w, R, w') = \log A(w, R, w') = \log \frac{p(w, w'|R)}{p(w|R)p(w'|R)} \tag{2.23}$$

Using all dependency triples extracted from the corpus by the parser as the data set, the likelihood ratio $A(w, R, w')$ is computed as

$$A(w, R, w') = \left(\frac{freq(w, R, w')}{freq(R)}\right)\left(\frac{freq(w, R)}{freq(R)}\right)^{-1}\left(\frac{freq(R, w')}{freq(R)}\right)^{-1} \tag{2.24}$$

In Lin (1998)'s convenient notation, frequency counts of the triple (w, R, w') are denoted as $||w, R, w'||$, with wild card expressions like $||w, R, *||$ used to denote the frequency counts of all triples that contain the specified elements of the triple. The

association score for the triple can then be written as:

$$I(w, R, w') = \log \frac{||w, R, w'||}{||*, R, *||} \cdot \frac{||*, R, *||}{||w, R, *||} \cdot \frac{||*, R, *||}{||*, R, w'||} = \log \frac{||w, R, w'|| \cdot ||*, R, *||}{||w, R, *|| \cdot ||*, R, w'||} \tag{2.25}$$

The similarity between words w_1 and w_2 is then computed as a sum of their association scores for the relation triples shared between the two words, divided by the sum of their association scores for all relations in which each of them occurs.

$$sim(w_1, w_2) = \frac{\sum_{T(w_1) \cap T(w_2)} I(w_1, R, w) + I(w_2, R, w)}{\sum_{T(w_1)} I(w_1, R, w) + \sum_{T(w_2)} I(w_2, R, w)} \tag{2.26}$$

where $T(a)$ represents the set of all relation/collocate pairs (R, w) such that $I(a, R, w) > 0$. In other words, unlike Hindle (1990), Lin (1998) discards entirely the shared "negative evidence" in distributional patterns, i.e. the relations/collocate pairs that occur with both target words less frequently that one would expect by assuming independence. Using this similarity measure, top-200 most similar words were extracted for each word in the corpus and evaluated against both WordNet and Roget's thesaurus. The measure was also used (1) to produce pairs of *reciprocal nearest neighbors* (Hindle, 1990) and (2) to experiment with sense induction. Sense induction involved building a similarity tree for the target word using its top-N most similar words. The target word was placed at the root. The other words, sorted by their respective similarity to the root, were then attached successively as offspring to whichever of the already existing nodes they happened to be most similar to. The subtrees of the resulting similarity tree represented different senses of the target word.

Using full context to group word occurrences into sense clusters

Schütze (1998) used bag-of-words co-occurrence features to create a word sense discrimination system based on "second order" context representations. In his system, each word was represented as a vector of frequencies with which other words were encountered with it (i.e. encountered within a certain window of the target word). The resulting word representation conflated senses of the target word. Such conflated *word vectors* were further used to create a *context vector* for each occurrence of the target word in the text. A *context vector* was obtained by summing (or taking an average in the normalized case) of the word vectors for all words making up that context (i.e. of all the words that occurred within a given window of the target word's occurrence). These context representations were further clustered into a pre-defined number of clusters, with each cluster representing a different sense of the target word.

A 50-word co-occurrence window was used in the experiments, with the dimensions of space for each target word chosen in one of two ways: (1) via a frequency cut-off,

using its 1000 most frequent neighbors, or (2) using a χ_2 criterion of dependence between the target word and the neighbor being selected as prospective "dimension". A combination of the EM algorithm and agglomerative clustering was used to cluster 2,000 randomly selected context vectors into senses. The cosine between vectors was used as a similarity metric. Each sense of the target word was then represented by the sense cluster centroid (the *sense vector*). These representations were further used to disambiguate unseen occurrences of the target word by selecting the *sense vector* closest to the corresponding context vector.

The results of summing the word vectors for the words occurring in a given context is that the features that co-occur with many words in the context will expand. What happens is essentially similar to topic detection. Different senses of the target word are effectively modeled as different "topics". Therefore, it's reasonable to expect that this method would do better with rather coarse topic-like ambiguities. Indeed, the evaluation was performed on the pseudowords – which would have precisely topic-like semantic distinctions – and on the ambiguous words with very coarse, almost homonymic sense differences[17]. Clearly, disambiguation frequently involves much more delicate distinctions. At the same time, the second order context representation provides a way to reasonably represent the full context of occurrence, rather than single syntactic relation as in the approaches above. And in many cases, a single grammatical relation/collocate pair is insufficient for disambiguation of the target word.

Sense detection via subtraction of tight cluster features

Pantel and Lin (Pantel and Lin, 2002; Pantel, 2003) used a full set of grammatical relations to induce an inventory of word senses for nouns, verbs, and adjectives. They define an MI-based association score between the word w and and a context of occurrence $c = (w', R)$, where w' denotes the word with which w occurs in relation R:

$$A(w, c) = \frac{p(w, c)}{p(w)p(c)} = \frac{||w, R, w'|| \cdot || * * * ||}{||w * *|| \cdot ||*, R, w'||} \tag{2.27}$$

A vector representation for the word w is constructed by multiplying each association score by a discounting factor:

$$\left(\frac{freq(c, w)}{freq(c, w) + 1}\right) \cdot \left(\frac{min(freq(c), freq(w))}{min(freq(c), freq(w)) + 1}\right) \tag{2.28}$$

- which decreases the association score when either the word itself or the context is infrequent, thereby adjusting for the fact that MI-based scores are higher for low

[17]Even so, better results are obtained for pseudo-words.

frequencies. Similarity between two words is computed as the *cosine* of their respective vectors.

A list of top-10 most similar words is then compiled for every word in the corpus[18]. The elements in each word's top-10 list (but not the word itself) are clustered using average-link clustering, and tightest and biggest cluster is stored[19]. The resulting list of stored clusters is sorted, with the precedence given to the tighter and bigger clusters. These clusters are then used to compile a set of *committees*. A *committee* is a tight cluster that represents a particular *sense* that exists in a language. A cluster is added to the set of *committees* if it falls far enough from every other cluster added the set so far. Similarity between clusters is computed as the cosine of cluster centroid vectors. Since the algorithm iterates over top-similar lists for all words in the language, the resulting set of committees should contain a committee for every *word sense* that exists in a language.

Every word in the language is then assigned to one or more *committees* as follows. Top-200 committees most similar to the word w are identified. If the committee cluster c most similar to w falls within a specified threshold, w is assigned to the sense represented by c. The features that c and w have in common are removed from the vector representation of w. The resulting residual representation of w is assigned to the sense represented by the next most similar committee cluster that falls within the same threshold[20]. The overlapping features are once again removed, and the procedure repeated, until no committee cluster in the remaining list falls close enough to the current representation of w. The results are thus dependent on two separate thresholds that must be set for (1) how far *committee* clusters should be from one another, (2) how close a given word should be to a committee for that committee to represent one of its senses[21].

Graph-based methods

Grammatical relation holding between words easily lend themselves to a graph representation. If words are represented as nodes, relations that are found to hold between them with sufficient frequency may be represented as (possibly, weighted) labeled directed edges. For example, Widdows and Dorow (2002) and Dorow and Widdows (2003), use the conjunctive relation between two nouns governed by "and"/"or" to build a co-occurrence graph. Node n is considered to have an edge leading to node

[18]In the interests of efficiency, the algorithm only considers the words that share high mutual information features with the original word.

[19]The cluster quality score for cluster c is computed as $|c| \cdot avgsim(c)$, reflecting a preference for clusters that are both tighter and bigger.

[20]The similarity is computed between w and the centroid of c. Note that if no cluster in the top-200 list for w falls within the threshold, w will not be assigned to any senses

[21]The presentation here simplifies slightly the actual algorithm used by Pantel and Lin (2002).

n' if n' occurs in the specified relation with n enough times to make it to n's top-N most frequent list. Note that n' may make the top-N of n, while n does not make the top-N list for n', so the result is a directed graph.

Node affinity score between a node u and a set of nodes A is defined as the percentage of that node's neighbors that are also neighbors of one of the nodes from A. A graph is built out of the words linked to the target word under a pre-set threshold. Senses are represented as connected components of the graph that remain in place when the node corresponding to the target word is removed from the graph.

2.2.4 Evaluation

Distributional similarity measures are typically evaluated either by comparing the output against a manually created resource, or indirectly via improving the performance of a particular NLP application. For example, overlap-based comparisons with WordNet synsets, Roget's thesaurus, and machine-readable dictionary definitions have repeatedly been used in evaluation (Grefenstette, 1994; Lin, 1998; Pantel, 2003). Distributionally similar words had also been used in query reformulation, judging collocational compositionality, and other tasks (Dagan, 2000; Weeds et al., 2004). If distributional similarity is used to resolve the sparsity problem, a frequency distribution predicted using distributional similarity may be evaluated directly by comparing it against the distributions in held-out data (e.g. Pereira et al., 1993). Weeds et al. (2004) proposed to shift the focus from such evaluations to the analysis of the linguistic and statistical properties of the obtained sets of distributionally similar words. However, they concentrate on relative word frequency of distributional neighbors, rather than on their semantic properties.

A manually constructed resource may be augmented with annotated corpora. For example, Pantel (2003) evaluates discovered word senses against WordNet classes. A WordNet class is subtree in WordNet hierarchy containing a synset and its hyponyms. Each WordNet class has a number associated with it, which is the probability of the corresponding sense (or any of its hyponym senses) occurring in the corpus. This probability is estimated using synset frequency counts from SemCor corpus. Probability of a higher-level class is therefore $\geq$ than the sum of probabilities of its hyponym classes. Valid WordNet classes against which evaluation is conducted are obtained by thresholding on the probability value. Words that have distinctive context features (with MI values higher than a certain threshold) are selected for the test set, and the test set is clustered using CBC.

Sense induction systems may also be evaluated directly against a gold standard of sense-annotated occurrences of a given target word. We discuss below some of the metrics that have been used for such evaluation.

Evaluation metrics for clustering solutions

A number of metrics have been proposed in the literature to evaluate the quality of a particular clustering solution against a gold standard (Amigó et al., 2008; Meila, 2003; Zhao and Karypis, 2004). Among them are set matching measures, measures based on edit distance, pairwise evaluation measures, measures based on mutual information, and some others.

We summarize some of the metrics proposed in literature below. We will use $C = \{c_i\}$ to refer to the set of clusters and $S = \{s_j\}$ to refer to the sense categories defined on the data set D, where data set $D = \{e\}$ is a set of elements (instances) to be clustered, and $n = |D|$. Note that a metric must support certain reasonable constraints, such as giving a lower score to the solution that merges two clusters that correspond to different senses, or unnecessarily splits a single sense.[22]

(1) *Set matching measures: Purity, Inverse Purity, F-score*

Under the set matching evaluation, a mapping is established between the induced clusters and the gold standard sense classes using precision, recall, or the F-measure.

Precision, recall, or the F-measure is computed for each cluster/sense class pair, and the pair that maximizes it for each cluster is used in the mapping. *Purity* is the weighted average of the precision values obtained for each cluster:

$$\textbf{Purity}(C, S) = \sum_i \frac{|c_i|}{n} \max_{i,j} \frac{|c_i \cap s_j|}{|c_i|} \tag{2.29}$$

Inverse Purity is the weighted average of recall values for each cluster:

$$\textbf{Inverse Purity}(C, S) = \sum_i \frac{|s_i|}{n} \max_{i,j} \frac{|c_i \cap s_j|}{|s_j|} \tag{2.30}$$

F-score matches each cluster to a sense class that maximizes the pairwise F-measure, i.e. the harmonic mean of precision and recall.

(2) *Pairwise evaluation measures*

Pairwise evaluation is based on checking whether pairs of elements that belong to the same cluster also belong to the same sense class.

Following Meila (2003), we adopt the following notation:

N_{11} denotes the number of pairs of elements that are both in the same cluster and in the same sense class;

[22] See, for example, Amigó et al. (2008) for similar considerations.

N_{00} denotes the number of pairs of elements that are neither in the same cluster, nor in the same sense class;
N_{10} denotes the number of pairs of elements that are in the same cluster, but not in the same sense class;
N_{01} denotes the number of pairs of elements that are not in the same cluster, but are in the same sense class.

Measures proposed by *Wallace*, *Fowkles and Mallows*, and *Rand* (cf. Meila, 2003), as well as the *Jaccard* coefficient, use pairwise evaluation. *Wallace*'s criteria compute the number of pairs correctly clustered together, divided by the total number of pairs clustered together (W_I), or by the total number of pairs that actually belong in the same sense category (W_{II}):

$$W_I(C,S) = \frac{N_{11}}{\sum_i |c_i|(|c_i|-1)/2} \tag{2.31}$$

$$W_{II}(C,S) = \frac{N_{11}}{\sum_j |s_j|(|s_j|-1)/2} \tag{2.32}$$

Fowkles and Mallows's criterion is the geometric mean of W_I and W_{II}.

Rand's criterion is computed as the total number of agreements between the clustering solution and the gold standard, i.e. the number of pairs that have been correctly classified together or correctly placed in different clusters, divided by the total number of pairs in the data set:

$$R(C,S) = \frac{N_{11}+N_{11}}{n(n-1)/2} \tag{2.33}$$

These criteria can be normalized by subtracting the agreement expected by chance for a given clustering and a given set of sense classes, and normalizing by the range. However, the useful range of the resulting measure varies depending on the particular clustering solution.

Jaccard index assesses the number of pairs correctly clustered together over the total number of pairs grouped together either by the clustering solution or under the gold standard:

$$J(C,S) = \frac{N_{11}}{N_{11}+N_{01}+N_{10}} \tag{2.34}$$

(3) *BCubed measures*

Amigó et al. (2008) define *BCubed Precision* an *Recall* to support a number of simplified constraints intended to give preference to clustering solutions with more homogeneous clusters and to penalize splitting a sense into two clusters. *BCubed Precision* an *Recall* effectively compute precision and recall *per element*, rather than *per pair*:

$$\textbf{BCubed Precision} = \frac{\sum_e \frac{|c_e \cap s_e|}{|c_e|}}{n} \tag{2.35}$$

$$\textbf{BCubed Recall} = \frac{\sum_e \frac{|c_e \cap s_e|}{|s_e|}}{n} \tag{2.36}$$

where $e \in D$ is an element of the data set, c_e is the cluster to which e belongs, and s_e is the sense category to which e belongs.

(4) *Entropy-based measures*

Entropy-related measures evaluate the overall quality of a clustering solution with respect to the gold standard sense classes.

Entropy of a clustering solution, as it has been used in the literature, evaluates how the sense classes are distributed with each derived cluster. It is computed as a weighted average of the entropy of the distribution of senses within each cluster:

$$\textbf{Entropy}(C, S) = -\sum_i \frac{|c_i|}{n} \sum_j \frac{|c_i \cap s_j|}{|c_i|} \log \frac{|c_i \cap s_j|}{|c_i|} \tag{2.37}$$

The *mutual information* of two variables defined by the clustering solution and the sense assignment $I(C, S)$ (cf. Meila, 2003) is defined as:

$$I(C, S) = \sum_{i,j} P(i, j) \log \frac{P(i, j)}{P(i)P(j)} \tag{2.38}$$

where $c_i \in C$ is a cluster from the clustering solution C, and $s_j \in S$ is a sense from the sense assignment S, and $P(i, j) = \frac{|c_i \cap s_j|}{n}$. The range for $I(C, S)$ depends on the entropy values of the two variables, $H(C)$ and $H(S)$:

$0 \leq I(C, S) \leq \min(H(C), H(S))$

Some other related measures have been proposed, for example, the *variant of information* measure (Meila, 2003), defined as $VI = H(C) + H(S) - 2I(C, S)$. This measure suffers from the same problem, i.e. its maximum depends on the respective entropy values.

CPA-Pattern → Segment verb-lit Segment | verb-lit Segment | Segment verb-lit | CPA-Pattern ';' Element
Segment → Element | Segment Segment | '' Segment '' | '(' Segment ')' | Segment '|' Segment
Element → literal | '[' Rstr ArgType ']' | '[' Rstr literal ']' | '[' Rstr ']' | '[' NO Cue ']' | '[' Cue ']'

Rstr → POS | Phrasal | Rstr '|' Rstr | epsilon
Cue → POS | Phrasal | AdvCue
AdvCue → ADV '[' AdvType ']'
AdvType → Manner | Dir | Location

Phrasal → OBJ | CLAUSE | VP | QUOTE
POS → ADJ | ADV | DET | POSDET | COREF POSDET | REFL-PRON | NEG |
MASS | PLURAL | V | INF | PREP | V-ING | CARD | QUANT | CONJ
ArgType → '[' SType ']' | '[' SType '=' SubtypeSpec ']' | ArgType '|' ArgType | '[' SType ArgIdx ']' |
'[' SType ArgIdx '=' SubtypeSpec ']'
SType → AdvType | TopType | Entity | Abstract | PhysObj | Institution | Asset | Location | Human |
Animate | Human Group | Substance | Unit of Measurement | Quality | Event |
State of Affairs | Process
SubtypeSpec → SubtypeSpec '|' SubtypeSpec | SubtypeSpec '&' SubtypeSpec | Role | Polarity | LSet
Role → Role | Role '|' Role | Beneficiary | Meronym | Agent | Payer
Polarity → Negative | Positive
LSet → Worker | Pilot | Musician | Competitor | Hospital | Injury | Ailment | Medicament |
Medical Procedure | Hour-Measure | Bargain | Clothing | BodyPart | Text | Sewage | Part |
Computer | Animal

ArgIdx → <number>
literal → word
CARD → <number>
POSDET → my | your | ...
QUANT → CARD | a lot | longer | more | many | ...

verb-lit → <verb-word-form>
word → <word>
NEG → not
INF → to

Table 2.1: CPA pattern grammar

$$Dice(A,B) = \frac{|A \cap B|}{\frac{1}{2}(|A|+|B|)}; \quad Dice^{\dagger}(\vec{X},\vec{Y}) = \frac{\sum_i \min(x_i,y_i)}{\frac{1}{2}\left(\sum_i x_i + \sum_i y_i\right)}$$

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|}; \quad Jaccard^{\dagger}(\vec{X},\vec{Y}) = \frac{\sum_i \min(x_i,y_i)}{\sum_i \max(x_i,y_i)}$$

$$\cos(\bar{X},\bar{Y}) = \frac{\vec{X}\cdot\vec{Y}}{|\vec{X}||\vec{Y}|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

$$Euclidean\text{-}Distance(\vec{X},\vec{Y}) = |\vec{X}-\vec{Y}| = \sqrt{\sum_i (x_i - y_i)^2}$$

$$L_1\ norm = \sum_i |x_i - y_i| = 2\left(1 - \sum_i \min(x_i, y_i)\right)$$

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

$$JS(p||q) = \frac{1}{2}\left[D(p||\tfrac{p+q}{2}) + D(q||\tfrac{p+q}{2})\right]$$

$$\alpha\text{-}skew(p,q) = D(p||\alpha \cdot q + (1-\alpha)\cdot p)$$

Table 2.2: Similarity measures

Chapter 3

Resolving Polysemy in Context

In this chapter, we examine the considerations that come into play when the ambiguity of the predicate is resolved based solely on the semantics of the arguments. We discuss the issues that have to be addressed by automatic algorithms that aim to group together the arguments that activate the same sense of a polysemous predicate. We look in more detail at different factors affecting sense assignment for both the verb and its arguments. We then discuss the difficulties that arise in designing sense inventories for cases when the semantics of the arguments is the major factor contributing to disambiguation.

3.1 Selection and Compositionality

Computational approaches to word sense disambiguation typically assume that each word in an utterance is assigned a sense from an inventory of senses. This is clearly a simplification of what actually happens when the meaning of a complex expression is computed. Consider a polysemous target predicate with certain semantic preferences. In a given argument position, different senses of that predicate will select for different semantic features. Thus, in (3.1a), the *pay* sense of *absorb* selects for ASSET in direct object position, while the *learn* sense selects for INFORMATION. Similarly, the *shoot* sense of *fire* in (3.1b) selects for PHYSOBJ = PROJECTILE, while the *dismiss* sense selects for PERSON.

(3.1) a. The customer will *absorb* this *cost*. [ASSET] (*pay*)
The customer will *absorb* this *information*. [INFORMATION] (*learn*)

b. The general *fired* four *lieutenant-colonels*. [PERSON] (*dismiss*)
The general *fired* four *rounds*. [PHYSOBJ] (*shoot*)

Selection is effectively a *bidirectional* process through which a particular interpretation is assigned both to the predicate and to its arguments. For example, in (3.2), the noun *rounds* is ambiguous between the TimePeriod and PhysObj. It activates the *shoot* sense of *fire*, and at the same time is itself disambiguated by the predicate.

(3.2) The general *fired* four rounds.

We will refer to such phenomena as *bidirectional selection*: the verb and its argument disambiguate each other directly, without any other elements of the context contributing to disambiguation. As such, it is related to, but distinct from *co-composition* (Pustejovsky, 1995) which occurs when both words in a dependency act as functors, creating a new, non-lexicalized sense for the composite expression. Bidirectional selection involves mutual disambiguation of two ambiguous words entering a dependency; it is not restricted to regular polysemies, and applies equally to any lexical ambiguity.[1]

The same sense of the predicate may be activated by a number of *semantically diverse arguments*. Such argument sets are frequently organized around a core of typical members that are a "good fit" with respect to semantic requirements of the corresponding sense of the target. The relevant semantic feature is prominent for them, while other, more peripheral members of the argument set, merely allow the relevant interpretation in context. Effectively, each sense of the target predicate may be seen to induce an ad-hoc semantic category in the relevant argument position. For example, consider two senses of the phrasal verb *take on* given in (3.3). Lexical items that occur in direct object position are given for each sense.

(3.3) a. Sense 1: *tackle an adversary*:
competition, rival, enemy, opponent, team, government, world.
b. Sense 2: *acquire a quality*:
shape, meaning, color, form, dimension, reality, significance, identity, appearance, characteristic, flavor.

The nouns in each of the above argument sets are quite distinct semantically, and yet activate the same sense of the predicate. The context provided by the predicate selects for a particular aspect of their sense, and the argument sets consist of a number of *core* elements for which it is a central component of their meaning and some *satellite* members for which the requisite component is peripheral. Thus, in the first argument set, the [+adversary] component is central for *enemy*, *rival*, *opponent* and *competition*, while *government* and *world* merely allow this interpretation due to animacy/agency.

[1]This phenomenon is not to be confused with *co-requirement* (Gamallo et al., 2005), which refers to the mutual restrictions imposed by the predicate and the argument on each other.

Core members of the argument set may be polysemous and require the *bidirectional selection* process in order to activate the appropriate sense of the predicate. But notice that the interpretive work that is done in (3.4a) and (3.4b), for example, is quite different.

(3.4) a. Are you willing to *take on* the competition?
b. Are you willing to *take on* the government?

While both words activate the same sense of *take on*, *competition* will merely be disambiguated between the EVENT reading and the ANIMATE, [+adversary] reading. For *government*, the [+adversary] reading will be coercively imposed by the predicate and is effectively accidental.

Another observation to make is that different aspects of meaning may be relevant for different dependencies the word enters into. For example, consider the use of the noun *opponent* with the verbs *take on* and *know* in (3.5a).

(3.5) a. It is much harder to *take on* the opponent you *know* personally.
b. It is much harder to *take on* the student you *know* personally.

FrameNet gives two senses for the verb *know*: (1) the *familiarity* sense (this is the sense in which you know people and places) and (2) the *awareness* sense (this is the sense in which you know propositional content). In this context, *opponent* activates the adversary reading for *take on* and the familiarity reading for *know*. While the first operation requires the [+adversary] component, the PERSON reading is sufficient for the second operation. Notice also that in (3.5b), the word *student* which is lacking the [+adversary] component, activates a different sense of *take on*.

3.1.1 Reusability of semantic features associated with argument sets

The same semantic component may be central to argument sets associated with different predicates. The question arises, to what extent are these argument sets "reusable" within a language. It becomes immediately clear, however, that each predicate imposes its own gradation with respect to prototypicality of elements of the argument set. As a result, even though basic semantic types such as PHYSOBJ, ANIMATE, EVENT, are used uniformly by many predicates, argument sets, while semantically similar, typically differ between predicates. For example, *fall* in the subject position and *cut* in the direct object position select for things that can be decreased:

(3.6) a. *cut (dobj)*: *reduce or lessen*
price, inflation, profits, cost, emission, spending, deficit, wages overhead, production, consumption, fees, staff
b. *fall (subj)*: *decrease*
price, inflation, profits, attendance, turnover, temperature, membership, import, demand, level

While there is a clear commonality between these argument sets, the overlap is only partial. To give another example, consider information-selecting predicates *explain (subj)*, *grasp (dobj)* and *know (dobj)*. The nouns *book* and *note* occur in the subject position of *explain*; *answer* occurs both as the subject of *explain* and direct object of *know*; however, *grasp* accepts neither of these nouns as direct object. Thus, the actual selectional behavior of the predicates does not seem to be well described in terms of a fixed set of types, which is what is typically assumed by many ontologies used in automatic WSD.

3.1.2 Selectors and sense separation

In case of homonymy, different senses of the predicate may select for semantic components that are quite distinct. In such cases, overall distributional similarity between arguments may be sufficient to group together the relevant lexical items. For example, *file* in the sense of *smooth* (e.g. *file nails, edges*, etc.) is easily distinguished from the cluster of senses related to *filing papers*. When the predicate's senses are related, this task is difficult even for a trained human eye.

Consider what happens if we need to determine which selectors are likely to activate what sense, keeping in mind that at least some of the verb's senses will be interrelated. Typically, corpus occurrences of a polysemous verb cluster into 2-10 groups, each roughly corresponding to a sense.[2] For each of these groups, one usually finds a lot of cases where sense distinctions are clear-cut and easily discernable. But whenever two senses are related, there are usually some *boundary cases* where it is not clear in which sense the predicate is used. Thus, in a given argument position, three kinds of selectors (i.e. NP heads) are possible:

(i) Good disambiguators: selectors that immediately pick one sense of the target. These selectors can be monosemous or polysemous themselves. When such selector is polysemous, its other sense(s) just never occur with the other sense of the target verb. Disambiguation is achieved through bidirectional selection, as in "fire four rounds" in (3.2).

[2]Light verbs have many more than that, but we will not consider them here.

(ii) Poor disambiguators: selectors that may be used with either sense and require more context to be disambiguated themselves (bidirectional selection doesn't work). For example, "assuming a position" may equally likely mean *taking on a post*, *adopting a particular bodily posture*, *occupying a certain point in space*, or *presupposing a certain mental attitude*, etc.

(iii) Boundary cases: the choice between two senses of the target is in fact impossible to make (i.e. the selector activates both senses at once).

For example, for the subject position with the verb *show* in (3.7), *survey* and *photo* are good disambiguators, while *graph* is a clear example of a boundary case.[3]

(3.7) a. The photo *shows* Sir John flanked by Lt Lampard.
(*pictorially represent*)
b. The survey *shows* signs of improvement in the second quarter.
(*demonstrate by evidence or argument*)
c. The graph *showed* an overall decrease in weight.
(both senses?)

Boundary cases are obviously identified as such only when there is enough good disambiguators for each of the related senses. For that reason, such cases are better construed as instances of simultaneous activation of both senses, rather than as evidence for overlapping sense definitions. We will refer to this phenomenon as *multiple selection*.

Interestingly, even syntactic pattern can not always overrule the interpretation intrinsic to some selectors. For example, in (3.8), it is virtually impossible to resolve *deny* between *refuse to grant* and *proclaim false*:

(3.8) a. Elders are often *denied* the status of adulthood
b. Philosophers have *denied* the autonomy to women

In (3.9), on the other hand, the selector itself is polysemous, with two interpretations available for it, and it needs to be disambiguated by context before it can activate the appropriate sense of the predicate.

(3.9) a. *deny* the traditional *view* (*proclaim false*)
b. *deny* the *view* of the ocean (*refuse to grant*)

[3]Apresjan (1973) gives a similar example of a boundary case between two senses of the verb *borot'sya* ("fight") in Russian: "fighting an opponent" vs. "fighting poverty, heresy". "Fighting heretics" is then seen as a clear boundary case.

In the following sections, we discuss how these considerations can be taken into account when defining a sense inventory.

3.2 Problems with Sense Inventories

3.2.1 Defining sense categories

As we have seen above, when the semantics of the arguments is the deciding factor in disambiguation, prototypicality – as a general principle of category organization – plays an important role in defining both the boundaries of senses and the corresponding argument groupings. For example, consider the verb *absorb*. One of its senses involves *absorbing a substance*. Typical members of the corresponding argument set would be actual substances, such as *oil, oxygen, water, air, salt*, etc. But *goodness, dirt, flavor, moisture* would also activate the same sense.

Each decision to split a sense and make another category is to a certain extent an arbitrary decision. Thus refining the senses for *absorb* further, one can separate *absorbing a substance* (*oil, oxygen, water, air, salt*) from *absorbing energy* (*radiation, heat, sound, energy*). The latter sense may or may not be separated from *absorbing impact* (*blow, shock, stress*). But it is a marked continuum, i.e. certain points in the continuum are more prominent, with necessity of a given concept reflected in the frequency of use.

3.2.2 Boundary cases

As we have pointed out above, when several senses are postulated based on the semantics of the arguments, there are almost always *boundary cases* that can be seen to belong to both categories. This affects both the sense inventory construction and the annotation. Consider, for example, two senses defined for the verb *launch* and the corresponding direct objects in (3.10):

(3.10) a. Sense 1: *Physically propel an object into the air or water*
missile, rocket, torpedo, satellite, shuttle, craft
b. Sense 2: *Begin or initiate an endeavor*
campaign, initiative, investigation, expedition, drive, competition, crusade, attack, assault, inquiry

The senses are quite distinct, yet examples like *launch a ship* clearly fall on the boundary: while *ships* are physical objects propelled into water, *launching a ship* can be virtually synonymous with *launching an expedition*.

To give another example, two senses of the verb *conclude* in (3.11) are linked to nominal complements and seem to be very clearly separated:

(3.11) a. Sense 1: *finish*
meeting, debate, investigation, visit, tour, discussion; letter, chapter, novel
b. Sense 2: *reach an agreement*
treaty, agreement, deal, contract, truce, alliance, ceasefire, sale

However, *conclude negotiations* is clearly a boundary case where both interpretations are equally possible (negotiations may be concluded without reaching an agreement). In fact, during the annotation we describe in Chapter 5, the two participating annotators chose different senses for this example:[4]

(3.12) We were able to operate under a lease agreement until purchase negotiations were concluded.
annoA: *finish*
annoB: *reach an agreement*

3.2.3 Regular semantic processes

In many cases, postulating a separate sense for a coherent set of nominal complements is not justified, as there are regular semantic processes that allow the complements to satisfy selectional requirements of the verb.

For example, the verb *conclude* in the *finish* sense accepts EVENT complements such as *visit*, *investigation*, etc. Nouns such as *letter*, *chapter*, *novel* in (3.11), while forming a semantically distinct cluster, activate the same sense as EVENT nouns. Such nouns are coerced into events corresponding to the activity that typically brings them about, that is, re-interpreted as events of writing (their Agentive quale, cf. Pustejovsky, 1995). A similar example is provided by the verb *deny*, which in the first sense (*state or maintain that something is untrue*) accepts PROPOSITION complements, as illustrated in (3.13):

(3.13) a. Sense 1: *state or maintain that something is untrue*
allegations, reports, rumour; significance, importance, difference; attack, assault, involvement
b. Sense 2: *refuse to grant something*
access, visa, approval, funding, license

[4]Semantic annotation task is described in Chapter 5. Here and elsewhere, we will refer to the two annotators that participated in the task as *annoA* and *annoB*.

EVENT nouns such as *attack* and *assault* activate the first sense by being coerced into a propositional reading, as do relational nouns such as *significance* and *importance*.

Similar regular semantic processes routinely account for variations within argument sets for other verbs, variations that do not warrant the definition of a separate sense. For example, consider another proposition-selecting verb, *believe*. Its direct object complements, given in (3.14), fall into several semantically distinct groups. Following the analysis given in Pustejovsky and Rumshisky (2008), the nouns in each group are still coerced to an interpretation of a proposition, although through different strategies. The nouns in (3.14a) either directly denote propositions (e.g., *lie*, *nonsense*) or are complex types that have an information component which can interpreted propositionally (e.g., *bible*, *polls*). The sources in (3.14b) are construed as denoting a proposition produced by (e.g., *woman*), or coming through (e.g., *ear*) the named source. Finally, the last set is licensed by negative polarity context, and is a state or event; e.g., "He couldn't believe his luck.").

(3.14) *believe.v*
object

a. PROPOSITION: lie, tale, nonsense, myth, opposite, truth, propaganda, gospel
b. SOURCE: woman, government, bible, polls, military; ear, eye
c. EVENT/STATE: luck, stupidity, hype, success

Similarly, following Pustejovsky and Rumshisky (2008), consider the argument sets for the direct object position of verbs such as *repair*, *fix* and *mend* which select for artifactual entities.

For these verbs, the same sense is activated by two kinds of lexical items: artifacts (i.e. man-made entities intended to serve a certain purpose, cf. Pustejovsky, 2001; Pustejovsky, 2006) and negative states representing the conditions of the artifactual entity, as in (3.15) and (3.17), and possibly also the general negative situation as in (3.16).

(3.15) *fix.v*
object

a. ARTIFACTUAL: pipe, car, alarm, bike, roof, boiler, lock, engine; heart; light, door, bulb
b. NEGATIVE STATE (condition on the artifact): leak, drip

(3.16) *repair.v*
object

a. ARTIFACTUAL: roof, fence, gutter, car, shoe, fencing, building, wall, pipe,

bridge, road; hernia, ligament
b. NEGATIVE STATE (condition on the artifact): damage, ravages, leak, crack, puncture, defect, fracture, pothole, injury
c. NEGATIVE STATE (general situation): rift, problem, fault

(3.17) *mend.v*
object
a. ARTIFACTUAL: fence, shoe, clothes, roof, car, air-conditioning, bridge clock, chair, wall, stocking, chain, boat, road, pipe
b. NEGATIVE STATE (condition on the artifact): puncture, damage, hole, tear

But in fact, the relevant sense of these verbs merely selects for a NEGATIVE STATE of an ARTIFACTUAL. When the negative relational state is realized, it can either take an artifactual as its object, or leave it implicitly assumed:

(3.18) a. *repair the puncture / leak*
b. *repair the puncture in the hose / leak in the faucet*

When the artifactual is realized, the negative state is left implicit by default.

(3.19) a. *repair the hose / faucet*
b. *repair the (puncture in) the hose / (leak in) the faucet*

The presence of distinct argument clusters in this case is therefore accounted for by the refinement in the semantic selectional specification for the verb. The same argument position merely gets filled by different semantic roles with respect to the relevant event.

3.2.4 Parallel sense distinctions

A very common problem with glossing a sense involves the situation where a sense inventory includes two senses one of which is an extension of the other. The derived sense may be related to the primary sense through metaphor, and this often results in the former taking on a semantically less specific interpretation. The problem with creating glosses in this situation is that the words used may have sense distinctions parallel to the ones in the target verb being described. This leaves the annotators free to choose either sense. This seems to be the case, for example, with OntoNotes sense inventory for *fire*, where *ignite or become ignited* is the gloss under which very divergent examples are grouped: *oil fired the furnace* (literal, primary sense) and *curiosity fired my imagination* (metaphoric extension). Clearly, annotators were

having a problem with this sense due to the fact that the verb *ignite* has sense distinctions which are based on the same metaphor (*fire* = *inspire*) and therefore are very similar to those of the verb *fire*.

3.2.5 Semantic underspecification

In case of semantic underspecification, annotators may be left free to choose the more generic sense, which contaminates the data set while not being reflected in the inter-annotator agreement values. Consider the sense inventory for *acquire* from the annotation task in Chapter 5. The gloss for usages such as *acquire a new customer* had to be very generic. We used the gloss "become associated with something, often newly brought into being". However, that led the annotators to overuse this gloss and select this sense in cases where a more specific gloss was more appropriate:

(3.20) By this treaty, Russia *acquired* a Black Sea *coastline*.
annoA: *become associated with something, often newly brought into being*
annoB: *become associated with something, ...*
correct: *purchase or become the owner of property*

For a more detailed analysis of this phenomenon, see Section 5.5.

3.3 Summary

In this chapter, we have examined the issues that arise when the senses of a polysemous verb are differentiated through the semantics of a particular argument. The main observation here is that the operation of sense assignment is bidirectional. The words that activate the same sense of the verb (when in a given argument position) may be quite distinct semantically. What unites them is that each of them either carries or is assigned a particular semantic interpretation, associated with the corresponding sense of the verb. As a consequence, identifying verbal sense distinctions that are linked to the semantics of the arguments involves recognizing the similarity of arguments in context.

In the following chapter, we use these considerations to design an automatic algorithm for inducing the verb senses associated with the semantics of the nouns in a particular argument position.

Chapter 4

Bipartite Contextualized Clustering

4.1 Preliminaries

4.1.1 Motivation

Given the considerations discussed in the previous chapter, it is clear that we need to be able to cluster together semantically diverse arguments that activate the same sense of the verb. In this chapter, we present an algorithm for solving this problem. We propose a solution based on using a contextualized representation of *selectional equivalents* of the target word and create a soft clustering assignment for selectors (i.e., NP heads) activating each sense.

First, let us look briefly at how the measures using the overall distributional similarity have been used to solve this problem. Approaches that use such similarity measures to group selectors according to the sense they activate achieve a certain degree of success. For example, Pereira et al. (1993) shows clusters for direct objects of the verb *fire* that pick out the "projectile" sense of *fire* from the "weapon" sense of *fire*, separating both from the "dismiss an employee" sense. An unpublished tool from Sketch Engine developers (Kilgarriff et al., 2004) uses a distributional similarity measure to cluster collocates of the target word in each grammatical relations. A cluster is created around each collocate with high association score with the target. For predicates, this effectively builds a simplified version of Lin's similarity tree (Lin, 1998) in every argument position.[1]

The resulting argument groupings are indeed often organized around a core set of semantically similar elements, with more peripheral members noticeably scattered. This simple observation served as the basis for a number of sense-induction algorithms.

[1]The Sketch Engine is reviewed in more detail in Appendix A.

For example, Pantel and Lin (2002) used tight-big clusters of low polysemy words to represent word senses, Velldal (2005) used for the same purpose a prototype vector based on a tight set of initial members of a fuzzy set.

However, a number of problems are difficult to resolve by using overall distributional similarity. For example, consider the verb *tackle*. Using the Sketch Engine's similarity tree constructed for the BNC data, it is possible to separate the sense corresponding to physically wrestling a person to the ground from the sense corresponding to considering an issue or a problem. However, in the BNC, there are very few instances of the first sense (*tackling a player, an intruder, a robber*). In order to group these together without destroying the grouping corresponding to the second sense, you have to fine-tune the clustering threshold quite carefully. Infrequent senses of the target verb often get lost in spurious near-synonym groupings. It is nearly impossible to identify instances where the argument's sense is modified through coercion. Often, the only way to create an argument cluster that picks a verb sense reliably is to tweak the clustering threshold manually.

Another observation about the distributional classes produced in automated thesaurus construction is that outside of the core set of near-synonyms (or antonyms), the resulting clusters often do not appear intuitive. For example, the following are the entries for the noun *rival* from Lin's Dependency-based Thesaurus (Lin, 2002) and Pantel's CBC output (Pantel, 2003) over the BNC:[2]

```
rival
competitor, opponent, challenger, candidate, contender, foe, ally,
"George W. Bush", Bush, Republican, front-runner, leader, Democrat,
gore, enemy, "vice president", supporter, company, McCain, "Bill Bradley",
nominee, Bradley, opposition, neighbor, adversary, conservative, politician
(Lin's Dependency-based Thesaurus)


rival
Nq750   leeway, free time, clout, spare time
Nq1151  impediment, obstacle, stumbling block, disincentive
Nq597   people, those, many, One
Nq749   minister, president, Prime Minister, government
(Pantel's CBC senses)
```

The reason some of these groupings seem unintuitive is that these words really do mean very different things *except* in a very specific context. The algorithm presented in this chapter was developed to address this issue directly by using the idea of contextualized similarity for sense induction. In the following sections, we give the motivation for our approach, present the proposed algorithm, and describe the system architecture and implementation.

[2]We give here only the top-ranking words from each thesaurus entry.

4.1.2 Contextualized Similarity

The goal of a similarity measure is to allow us to tell automatically whether one word is "like" the other. But whether one word is like the other may vary, depending on the particular task. If our task is to determine the meaning of a predicate by looking at its arguments, two words in the same argument position will be "like" each other only if they pick the same sense of the predicate. We can capture this intuition by defining a measure aimed to assess *contextualized similarity*, i.e. similarity between two lexical items with respect to a particular context.

We adopt a context representation based on the notion of grammatical relation as it is used in distributional similarity literature (see, e.g. Lin, 1998; Hindle, 1990). An instance of a grammatical relation is a tuple (w_1, R, w_2), where R denotes the type of grammatical dependency between the words w_1 and w_2. A context is a set of such tuples, as extracted from a single instance of occurrence of the target word. For example:

(4.1) sentence: Their life *took on a different meaning.*
context(*meaning*): {(*take on*, obj, *meaning*), (*meaning*, modifier, *different*)}
context(*take on*): {(*take on*, obj, *meaning*), (*take on*, subj, *life*)}

In the following discussion, we will use the term *context* to refer to a singleton, i.e. a single populated syntactic relation.

At its most basic, distributional similarity between frequency profiles of two words should reflect to what extent the contexts in which the two words occur overlap. Similarity between two words may be expressed as the frequency of their occurrence in identical contexts, relative to the average of their overall frequencies. Since the two words may have very different corpus frequencies, some normalization is also typically used. The result is a function of tuple frequency, typically referred to as the *weighting* or the *association score* between the word and the context attribute[3].

Defined in this manner, distributional similarity will be high for lexical items whose *overall* distributional profiles are similar. This will be the case for words which are semantically very close in their dominant, most frequent sense. Or, in a less likely case, it may be that most of their senses are similar, and have similar relative frequencies. In case of selectors that activate the same sense of a polysemous word, high similarity values may be obtained for the elements of semantically uniform core of the selector set (when such a core is present). Polysemous core elements for which the relevant sense is not dominant, as well as peripheral elements of the selector set, will slip through the cracks.

Hindle (1990) remarks that while one can *have* and *sell* both *beer* and *wine*, it's

[3]See, for example, Curran, 2004 for a survey of different weighting schemes.

the fact that you can *drink* both of them that makes them semantically close. In other words, when computing semantic similarity based on distributional behavior, some contexts are, to quote Orwell, "more equal than others". The reason we know that two words are used similarly in a given context is that there is a number of other contexts in language where they are used in the same way. Such *licensing contexts* license the use of these lexical items with the same sense of the target word.

Consider, for example, selectors for the two senses of *take on* in (3.3): *competition, rival, opponent, government* vs. *shape, meaning, dimension, significance*. Table 4.1 shows some of the contexts in which these selectors occur.[4] The fact that both *significance* and *shape* occur as direct objects of such verbs as *retain*, *obscure*, and *acquire* allows them to activate the *acquire a quality* interpretation for *take on*. Note that licensing contexts do not need to be syntactically parallel to the target context. So (*struggle*, pp_against) may select for the same semantic property as the *tackle an adversary* sense of (*take on*, obj).

When computing contextualized similarity for two selectors, we would like to give higher weights to the terms that correspond to the licensing contexts. Consider, for example, using the contexts shown in Table 4.1 to compute similarity between *competition* and *government* as direct objects of *take on*.[5] Their association scores with contexts similar to the target context must have higher weight than their association scores with non-similar contexts, i.e. (*threaten*, obj), (*confront*, obj) and (*struggle*, pp_against) should carry a higher weight than (*prize*, n-modifier) or (*the*, det). When both selectors occur in an unrelated context, the latter may in fact activate a completely different reading for each of them. For example, in the phrase "competition prize" *competition* is interpreted as an EVENT, and not as ANIMATE, +adversary. Consequently, the fact that both *government* and *competition* occur as nominal modifiers of *prize* should not be regarded as evidence of their similarity as direct objects of *take on*.

Computing similarity between contexts thus poses a separate problem. It is clearly incorrect to use overall distributional similarity between context-defining words to determine how close two contexts are. In order to be considered similar, two contexts must be similar with respect to their selectional properties. We introduce the notion of *selectional equivalence* below as a way of addressing this problem.

4.1.3 Selectional Equivalence

Selectional equivalence is defined for two verbs with respect to a particular argument position, and a particular sense for each verb. If nouns can be organized into lexical

[4] Association scores shown in the table are conditional probabilities $P(\text{selector}|\text{context})$.

[5] pp_against is a relation between the governing verb and the head of a prepositional phrase introduced by *against*; n_modifier is a relation between a noun and a nominal modifier.

target context: (take on, obj)

phrase	context	selectors, P(selector|context)			
		significance	shape	competition	government
retain __	(retain, obj)	.0030	.0030	.0000	.0000
obscure __	(obscure, obj)	.0016	.0043	.0000	.0000
acquire __	(acquire, obj)	.0043	.0006	.0000	.0000
threaten __	(threaten, obj)	.0000	.0008	.0008	.0057
confront __	(confront, obj)	.0000	.0000	.0009	.0104
struggle against __	(struggle, pp_against)	.0000	.0000	.0008	.0089
__ prize	(prize, n_modifier)	.0000	.0000	.0069	.0005
the __	(the, det)	.0000	.0000	.0000	.0000

Table 4.1: Sample licensing contexts for selectors of *take on*.

sets sharing a semantic feature, verbs can be organized into selectional equivalence sets, with arguments sharing a semantic feature.

A lexical item w_1 is a *selectional equivalent* of lexical item w_2 with respect to grammatical relation R, if in the argument position defined by R, one of the senses of w_1 selects for the same aspect of meaning as one of the senses of w_2. Such selectional equivalence can also apply to two lexical items w_1 an w_2 with distinct relations R_1 and R_2. Selectional equivalents do not need to be synonyms or antonyms of each other. Their equivalence is only in terms of the aspect of meaning they select. They are *contextual synonyms* of each other. Verbs that are selectionally equivalent to one of the senses of the target verb effectively form a subset of all licensing contexts for that sense.

If we can measure how close two contexts are with respect to the target context, selectional equivalents can be grouped into clusters representing different senses of the target verb. Resulting clusters can then used to determine how likely each selector is to be associated with that sense. We outline this procedure below in Section 4.2.1. Clusters of selectional equivalents obtained for selected senses of *take on*, *launch*, and *deny* are shown in (4.2).

(4.2) a. *take on* (*acquire a quality*)
acquire, obscure, assume, retain, possess

b. *launch* (*begin*)
organize, mastermind, spearhead, orchestrate, mount; commence, initiate, instigate, intensify, complete, undertake

c. *deny* (*proclaim false*)
confirm, disclose, conceal, reveal, uncover, corroborate, rebut, substantiate, disprove, refute, contradict, retract, furnish, gather, cite, collate, produce,

detail, present, summarize, suppress, publicize

d. *deny* (*refuse to grant*)
refuse, grant, revoke, obtain, withhold

Selectional equivalence thus implies a specific kind of semantic similarity, which overlaps only partially with what manually constructed resources typically aim to capture. In FrameNet, for example, selectionally equivalent verbs may belong to the same frame, or to the frames related through some frame-to-frame relation, such as frame inheritance or the Using relation. This is reasonable, since one would expect semantically uniform core elements to be similar when the verbs that operate on them are from the same situational frame. For example, *deny* and *confirm* in (4.2c) both evoke the same *Statement* frame; *disclose* and *reveal* evoke the frame which inherits from *Statement*. On the other hand, pairs such as *obscure* and *assume* in (4.2a) are not likely to evoke related frames. The same partial overlap can be observed with Levin classes and WordNet categories.

In order to obtain clusters of selectional equivalents for each sense of the target verb, we need to be able to measure to what extent two verb senses share selectional properties. This measure of selectional equivalence effectively mirrors contextualized similarity as defined for selectors. The idea is to take all selectors that occur in the specified argument position with the target verb, identify the verbs that occur with these selectors, and cluster them according to the sense of the target with which they share selectional properties. Our model involves the assumption that two verbs tend to be selectionally close with respect to just one of their senses. Similarity between two verbs is estimated based on selectors that, for each of them, consistently activate the sense which is selectionally equivalent to one of the target's senses.

In the next section, we outline the overall architecture of the algorithm and discuss in more detail the choice of reliable selectors. We then look at some results of the similarity computation based on the obtained selector lists.

4.2 System Architecture

Consider a bipartite graph where one set of vertices corresponds to headwords and the other to dependents, under a relation R. Each relation can be viewed as a function mapping from headwords to dependents.[6] The relation is defined by a set of tuples (w, R, w'), where w is the head, and w' is the dependent. The inverse of each relation is then a set of tuples (w', R^{-1}, w).

[6]This graph representation is similar to the one used in literature more commonly for symmetric relations such as conjunction or apposition (Widdows and Dorow, 2002) or co-occurrence within a window (Agirre et al., 2006).

Our system produces clusters of *selectional equivalents* for each sense of the target word, which induces the clustering of selector contexts according to the sense of the target word which they activate.

4.2.1 Algorithm Description

A corpus is tokenized, POS-tagged, and parsed. Grammatical relation tuples are extracted for all lemmas. For each target word t and relation R, we execute the following steps: (1) establish the set of words to be clustered, i.e. identify potential candidates for selectional equivalency for all senses of the target word, (2) identify reliable selectors for each potential selectional equivalent, and (3) produce clusters of selectional equivalents.

We give a detailed description of each step below.

4.2.1.1 Establishing the set of words to be clustered

This preliminary step is accomplished as follows:

(1) Identify the set of selectors with which the target word occurs in relation R. For example, for $t = \textit{acquire}$, $R = \mathsf{obj}$, this gives the set of nouns that occur in direct object position with *acquire*.

(2) Take the inverse image of that set under the R^{-1} relation.[7] In the example above, this operation produces a set of verbs which occur with direct objects of *acquire*, i.e. a set of candidates for selectional equivalency for different senses of the verb *acquire*.

The resulting set of words is sorted according to the number of the target's selectors with which they co-occur, with words that occur with less than 2 distinct selectors thrown out. For efficiency, we restrict the number of elements to be clustered to 4000, selecting the words that co-occur with a higher number of target's selectors.

4.2.1.2 Identifying reliable selectors

Since every word w in the resulting set occurs with some of the same selectors as the target t, it could potentially be selectionally equivalent to one of the target's senses. We need to identify selectors that for both t and w behave in the following manner: (1) activate the appropriate sense (2) are good disambiguators, i.e. the ones that activate only one sense and are not likely to occur with the other senses. Such selectors can be polysemous themselves, but merely always occur in the same sense

[7] We discard the cases that occur together only once.

when combining with t or w. If a selector occurs frequently with both t and w, several explanations are possible:

(i) A selector activates the appropriate sense for both t and w, and that sense is fairly frequent for both words:

a. *take on*/*acquire* a new *importance*

(ii) (*Parallel Sense Distinctions.*) If the verbs have more than one selectionally equivalent sense, a selector could activate the wrong pair of senses:

a. *acquire*/*possess* a new *significance* (QUALITY)
b. *acquire*/*possess* a powerful *weapon* (POSSESSION)

(iii) (*Selector Polysemy.*) Different senses of that selector may activate unrelated interpretations for the two verbs:

a. *take on* a greater *share* of the load
b. *acquire* the *shares* of the company

In our model, we make an assumption that the first case is the dominant one, while the other two cases are much more rare. Under such conditions, selectors that are strongly associated with both t and w must be the ones that pick the corresponding sense for each of them.[8]

For every word in the set of candidates for selectional equivalence, we obtain a set of reliable selectors as follows:

1. For each selector s that occurs both with t and w, compute association score $assoc_R(s, w)$ and $assoc_R(s, t)$.

2. Combine the two association scores using a combiner function $\psi(assoc_R(s, w), assoc_R(s, t))$ and choose the top-k selectors that maximize it.

Each w is then represented as a k-dimensional vector $\overline{w} = \langle f(s) \rangle$, where $f(s)$ is selector scoring function that determines the value for each selector based on its association scores.

For example, consider the verbs *acquire* and *lack* which are selectionally equivalent with respect to one of the senses of *acquire* (*take on a certain characteristic*). We would like for $assoc_{\mathsf{obj}}$(*importance-n,acquire-v*) and $assoc_{\mathsf{obj}}$(*importance-n,lack-v*) to

[8] A selector that is strongly associated with both t and w must occur "frequently enough" with each of them. Ideally, the frequency of distribution on the senses for w and t must be taken into account, since the relevant sense may be much more prominent for one word than for the other.

produce a combined value that is high enough to allow *importance* to be identified as a reliable selector.

We tested several system configurations which varied with respect to the association score used, the combiner function ψ and the selector scoring function $f(s)$. We summarize different configurations and explain the motivation for different configuration choices in Section 4.2.2.

4.2.1.3 Producing clusters of selectional equivalents

We use *group-average agglomerative clustering* to produce clusters of selectional equivalents $C_i = \{w\}$ for each sense of the target word, with each w represented as a k-dimensional vector. We remove the contribution of unreliable selectors by using low values for k, such as $k = 15$, with the cutoff point determined empirically.

In group-average agglomerative clustering, a similarity matrix is initially computed for all element pairs. At the start, each cluster contains just one element. Similarity between two clusters is computed as an average of pairwise similarity values between the elements of two clusters.[9]

Computing Similarity Similarity for two elements w_1 and w_2 is computed as the numeric equivalent of set intersect (i.e. sum of minimums[10]) for the scores assigned to the top-k reliable selectors chosen for w_1 and w_2.

$$csim_k(w_1, w_2, (t, R)) = \sum_{s \in S_1 \cap S_2} \min(f_1(s), f_2(s)) \tag{4.3}$$

where t is the target word, R is the grammatical relation, S_i is the set of top-k reliable selectors that pick the same sense of w_i and t, and $f_i(s)$ is the score assigned to selector s in the vector representation of w_i.

We discuss this choice of the similarity measure in Section 4.2.2.

Intra- and Inter-cluster APS During the agglomerative group-average clustering, similarity for all element pairs is used to compute *Average Pairwise Similarity* (APS) between every two clusters. This is accomplished by keeping track, for every pair of merged clusters (C_i, C_j), of the sum of pairwise similarities for all element pairs (w, w'), where $w \in C_i$ and $w' \in C_i$. This sum is divided by the total number of such pairs to obtain *inter-cluster APS* for two clusters:

$$\textit{inter-cluster APS}(C_i, C_j) = \frac{\sum_{w \in C_i, w' \in C_j} csim(w, w')}{|C_i| \cdot |C_j|} \tag{4.4}$$

[9] For more detail, see, for example, Manning and Schütze (1999).

[10] See Section 2.2.2

All cluster pairs are kept on a sorted queue, and the pair maximizing this value is merged at the next step. During each merge, we also keep track of *intra-cluster APS* for the resulting cluster, i.e. the average pairwise similarity between the elements of one cluster. As we proceed from the bottom of the dendrogram up, *intra-cluster APS* for the clusters decreases. We compute the percent decrease in *intra-cluster APS* (i.e the derivative) for every cluster merge point.

Ranked selectors lists As the dendrogram is built, we keep a list of selectors for each node in the tree. When two clusters are merged, a union of their selector lists is computed. Each selector is assigned a score that is a weighted average of its scores in the merged clusters (weighted by the number of elements in the cluster). The resulting selector list is sorted by the scores computed for each selector.

The way ranked selector lists are produced is illustrated in Figure 4.1 which shows an excerpt of a merge trace for the target context $(t, R) = (\textit{acquire-v}, \mathsf{obj})$. For each new cluster, the trace shows the cluster id, the ids of the two merged clusters, the *inter-cluster APS*, the *intra-cluster APS* of the resulting cluster, the elements of each of the merged clusters, and the selector list for the resulting cluster with the association scores.[11] The trace in Figure 4.1 shows two clusters being merged, [*emphasise-v*] and [*stress-v underline-v*]. The resulting cluster, [*emphasise-v stress-v underline-v*], contains selectional equivalents for one of the senses of *acquire* (*take on certain characteristics*, cf. p. 74). Each of its selectors has a score that is a weighted average of the scores of the two merged clusters.

Soft clustering for selectors While the resulting dendrogram establishes *hard clustering* for the target's selectional equivalents, we also obtain *soft clustering* for selectors.[12] The *ranked selector list* obtained for each cluster effectively provides soft cluster assignment for selectors. This is consistent with the fact that each selector may activate more than one sense of the target. These selector lists are used to optimize cluster choice in word sense induction task, cf. Chapter 6. However, since we are using group-average clustering, they are *not* used to compute similarity between the pairs of clusters.

[11] For the data set used in Figure 4.1, selectors were chosen using a harmonic mean of $assoc_R(s, w)$ and $assoc_R(s, t)$, with association score being the mutual information between selector s and potential selectional equivalent w.

[12] In *hard clustering*, each element assigned to one cluster. In *soft* or *probabilistic clustering*, elements may be assigned to multiple clusters, with an association score given for each cluster. For more detail, see Manning and Schütze (1999), Hastie et al. (2001), and others.

```
Cluster 2234=564+667 (45.351/45.351) [stress-v] [underline-v]
<pre-eminence-n:8.91 distinctiveness-n:8.77 significance-n:7.47 credentials-n:7.26
importance-n:7.20 dimension-n:6.99 salience-n:5.15 respectability-n:4.17
reputation-n:3.83 humility-n:3.72 individuality-n:3.70 liturgy-n:3.66 normality-n:3.57
urgency-n:3.43 liking-n:3.39 gloss-n:3.37 fascination-n:3.33 status-n:3.29
elegance-n:3.29 hollow-n:3.25 competence-n:3.25 orientation-n:3.24 sensitivity-n:3.16
willingness-n:3.14>

Cluster 747 [emphasise-v]
<distinctiveness-n:8.55 stigma-n:8.25 longevity-n:8.11 tan-n:7.89 individuality-n:7.66
credentials-n:7.66 legitimacy-n:7.32 significance-n:7.25 importance-n:7.17
hollow-n:6.94 reputation-n:6.81 attribute-n:6.69 trait-n:6.50 relevance-n:6.41
status-n:6.32>

Cluster 2239=747+2234 (43.648/44.215) [emphasise-v] [stress-v underline-v]
<distinctiveness-n:8.70 significance-n:7.40 credentials-n:7.39 importance-n:7.19
pre-eminence-n:5.94 individuality-n:5.02 reputation-n:4.82 dimension-n:4.66
hollow-n:4.48 status-n:4.30 salience-n:3.43 respectability-n:2.78 stigma-n:2.75
longevity-n:2.70 tan-n:2.63 humility-n:2.48 legitimacy-n:2.44 liturgy-n:2.44
normality-n:2.38 urgency-n:2.29 liking-n:2.26 gloss-n:2.24 attribute-n:2.23
fascination-n:2.22 elegance-n:2.19 trait-n:2.17 competence-n:2.16 orientation-n:2.16
relevance-n:2.14 sensitivity-n:2.11 willingness-n:2.09>
```

Figure 4.1: Merging ranked selector lists for (*acquire*, obj).

4.2.2 System Configurations

Several configurations of the system were implemented. The configurations vary with respect to the association score used, the method used to pick the top-k selectors, and the selector scoring function.

We used three types of association scores: conditional probability $P(s|Rw)$, pointwise mutual information mi, and mi multiplied by a log factor of the tuple count $freq(s, R, w)$. Selector scoring function $f(s)$ was one of the following: (1) the association score $assoc_R(s, w)$ itself, (2) the product of the selector's association scores with w and t, or (3) the harmonic mean of the two association scores. The combiner function ψ was either the geometric or the harmonic mean of the selector's association scores with w and t. In case the selector scoring function used a combination of the association scores with t and w, the same function was used to sort selectors.

Resulting 12 configurations are summarized in Table 4.2. We assume the following notation:

$assoc_w = assoc_R(s, w)$ is the association score between s and w

$hmean(a, b) = \frac{2ab}{(a+b)}$ is a harmonic mean

$LF = \log(freq(s, R, w))$

$mi(s, Rw) = \log \frac{P(s,R,w)}{P(s)P(R,w)}$ is pointwise mutual information

Probability values computed as follows:

$$P(s|Rw) = \frac{freq(s,R,w)}{freq(*,R,w)}$$

$$P(s) = \frac{freq(s,*,*)}{freq(*,*,*)}$$

$$P(R,w) = \frac{freq(*,R,w)}{freq(*,*,*)}$$

$$P(s,R,w) = \frac{freq(s,R,w)}{freq(*,*,*)}$$

where $freq(s,R,w)$ is the number of tuples extracted for grammatical relation R, headword w, and dependent s; $freq(*,R,w)$ is the number of tuples extracted for R and w occurring with any dependent; $freq(s,*,*)$ is the number of dependency tuples extracted for s; and $freq(*,*,*)$ is the total number of tuples extracted from the corpus.[13]

$assoc_w$	$f(s)$	$\psi(assoc_w,\ assoc_t)$	**configuration**
$P(s\|Rw)$	$assoc_w$	$assoc_w \cdot assoc_t$	CP-PROD
$P(s\|Rw)$	$assoc_w$	$hmean(assoc_w, assoc_t)$	CP-HMEAN
$P(s\|Rw)$	$assoc_w \cdot assoc_t$	$assoc_w \cdot assoc_t$	CP-PROD-PROD
$P(s\|Rw)$	$hmean(assoc_w, assoc_t)$	$hmean(assoc_w, assoc_t)$	CP-HMEAN-HMEAN
$mi(s,Rw)$	$assoc_w$	$assoc_w \cdot assoc_t$	MI-PROD
$mi(s,Rw)$	$assoc_w$	$hmean(assoc_w, assoc_t)$	MI-HMEAN
$mi(s,Rw)$	$assoc_w \cdot assoc_t$	$assoc_w \cdot assoc_t$	MI-PROD-PROD
$mi(s,Rw)$	$hmean(assoc_w, assoc_t)$	$hmean(assoc_w, assoc_t)$	MI-HMEAN-HMEAN
$mi(s,Rw) \cdot LF$	$assoc_w$	$assoc_w \cdot assoc_t$	MI-FACT-PROD
$mi(s,Rw) \cdot LF$	$assoc_w$	$hmean(assoc_w, assoc_t)$	MI-FACT-HMEAN
$mi(s,Rw) \cdot LF$	$assoc_w \cdot assoc_t$	$assoc_w \cdot assoc_t$	MI-FACT-PROD-PROD
$mi(s,Rw) \cdot LF$	$hmean(assoc_w, assoc_t)$	$hmean(assoc_w, assoc_t)$	MI-FACT-HMEAN-HMEAN

Table 4.2: System configurations

The accuracy of the resulting selector/sense assignment clearly depends on the success of each stage of the algorithm. We illustrate below the outcome of different stages of the algorithm using the target context $(t,R) = (deny, \mathsf{obj})$

Association scores Conditional probability $P(s|Rw)$ gives equal weight to every instance of the same selector s occurring with w and t, regardless of how frequent s itself is. The frequency of occurrence of a given selector in each context (i.e., (R,w) or (R,t)) is only normalized by the frequency of that context. Dividing the

[13]The notation here is similar to the one used by Lin (1998).

resulting association score by the frequency count for s gives more weight to the less frequent selectors. Hence, using mutual information as the association score results in selector lists being comprised by the nouns that are less frequent but perhaps more "characteristic" of the particular sense they select. The downside of this situation is that selector lists for selectional equivalents of the same sense may have fewer elements in common, thus potentially making the similarity computation less reliable. Log factor de-emphasizes selectors with low occurrence counts relative to the more frequent ones, but does so not as strongly as it would be done by simply removing the normalization by selector frequency.

Combiner functions Combiner function provides the score that has to be maximized by the selectors chosen to represent each element w. This is the step that insures that the resulting vector is contextualized with respect to the target context. Selectors that are picked must strongly associate with both t and w .

Using the product (or, equivalently, the geometric mean) of the two association scores as the combiner function to sort selectors for each w induces a sorting order with the sequence of equivalence classes located along the hyperbolic curves. If the relevant sense is infrequent for the target, but predominant for w, the combined score would still be fairly high.

We would like to avoid producing a high combined score in a situation when one of the association scores is much smaller than the other. The geometric mean gives equal weight to both values (i.e., increasing the smaller value by a certain factor increases the mean by the same factor as would increasing the larger value). Harmonic mean gives more weight to the increase in the smaller value, giving a preference to selectors that have similar association scores with both w and t.

Table 4.3 shows selectors chosen for the direct object position of two selectional equivalents of the verb *deny*, namely, *grant*, and *confirm*.[14] Selector quality for each pair of contexts is estimated as the geometric mean of conditional probabilities. Good disambiguators are shown in italic. Notice that *confirm-v* and *grant-v* are selectional equivalents of two different senses of *deny-v*. Selector sets chosen for the verb pairs *deny-v / confirm-v* and *deny-v / grant-v* reflect this distinction. That is, *report*, *allegation*, *importance*, etc., while quite diverse semantically, are typically *denied* in the same sense in which they can be *confirmed*, i.e. as PROPOSITIONS. Likewise, *access*, *approval*, *request*, etc., are *denied* in the same sense in which they can be *granted*.

Figure 4.2 illustrates how selectors are picked, with association scores for the target $assoc_R(s,t)$ along the x-axis, and association scores for the selectional equivalent $assoc_R(s,w)$ along the y-axis. Good disambiguators are depicted in green. Clearly,

[14]The table shows the data from the British National Corpus (BNC, 2000), with RASP (Briscoe and Carroll, 2002) used to extract grammatical relations.

automatically identifying all good disambiguators is not feasible. Our goal is to choose enough selectors correctly so that selectional equivalents of the same sense can be grouped together.

	deny-v		*grant-v*			*deny-v*		*confirm-v*	
	count	$P(n\|Rv)$	count	$P(n\|Rv)$		count	$P(n\|Rv)$	count	$P(n\|Rv)$
access	110	.0273	56	.0129	*report*	103	.0256	62	.0159
right	57	.0141	46	.0108	*existence*	92	.0228	32	.0082
approval	46	.0114	57	.0132	*claim*	77	.0191	17	.0043
permission	9	.0022	228	.0528	*allegation*	99	.0246	7	.0018
rights	23	.0057	63	.0145	*view*	8	.0019	86	.0221
status	15	.0037	74	.0171	*importance*	32	.0079	18	.0046
charge	184	.0457	5	.0011	*fact*	20	.0049	23	.0059
power	9	.0022	60	.0139	*involvement*	63	.0156	6	.0015
request	15	.0037	36	.0083	*charge*	184	.0457	2	.0005
license	2	.0049	254	.0588	right	57	.0141	6	.0015

Table 4.3: Top 10 selectors chosen for the verb pairs *deny-v/grant-v* and *deny-v/confirm-v* in direct object position. Correctly chosen selectors are italicized.

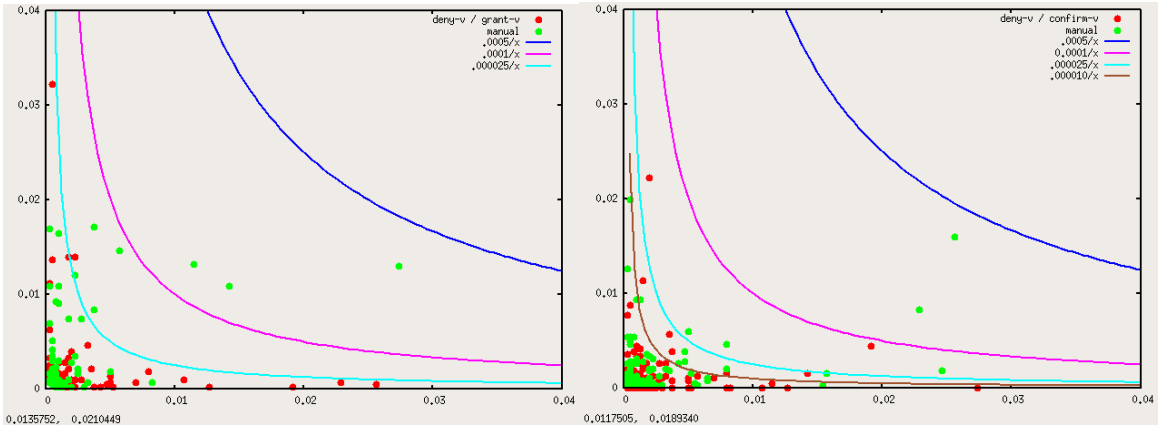

Figure 4.2: Choosing *good selectors* for the verb pairs *deny-v/grant-v* (left) and *deny-v/confirm-v* (right), representing two different senses of *deny-v*.

Selector scoring function Selector lists are chosen to contextualize the representation of each potential selectional equivalent to the target context. Using a combination of the association scores $assoc_R(s, w)$ and $assoc_R(s, t)$ takes contextualization of the resulting representation one step further, making more prominent the scores for selectors that are strongly associated with the target.

Similarity measure Group-average clustering produces relatively compact clusters that are relatively far apart. In computing pairwise similarity we do not normalize the sum of minima either by the size of the union, or by the average size of each set S_i (unlike the standard numerical extensions of Jaccard and Dice measures).

For the MI and MI-FACTOR scoring schemes, normalization is effectively incorporated into the association score. For the CP scheme, in order to avoid obtaining high similarity scores for high-frequency words among potential selectional equivalents, we need to avoid normalization. For example, you can *see* and *describe* most of the things you can *take on*, but that does not make them good selectional equivalents for either of the senses of *take on*. Effectively, these are promiscuous predicates that occur frequently with all selectors, including reliable selectors for each of the target word's senses. Conditional probabilities for their selectors, however, are low due to their high frequencies. Normalizing the sum of minima by the sum of maxima, as in Jaccard, for example, would bring up the similarity value for high-frequency pairs such as *see* and *describe*. Without such normalization, both words in such pairs have equally low values for all nouns in their respective selector lists, which leads to a low similarity score.

	refuse	confirm	contradict
report	0.000	0.018	0.006
claim	0.004	0.007	0.019
story	0.000	0.004	0.004
view	0.000	0.023	0.032
allegation	0.000	0.001	0.002
suggestion	0.000	0.002	0.006

	confirm	grant	refuse
access	0.000	0.013	0.014
rights	0.001	0.015	0.002
permission	0.000	0.053	0.066
request	0.001	0.008	0.034
relief	0.000	0.012	0.009
application	0.001	0.014	0.054
bail	0.000	0.016	0.011

Table 4.4: Similarity computation for selectional equivalents of two senses of *deny*. Association scores $P(n|Rv)$ for the intersection of top-k selector lists are shown for: A. (left) *confirm* and *contradict*, as compared with *refuse*. B. (right) *grant* and *refuse*, as compared with *confirm*.

There are inevitable misfires in the obtained selector lists. However, in order to compute the similarity value, we use the intersection of selector lists (cf. Eq. 4.3). For selectional equivalents of the same sense, this discards most of the spurious selectors chosen for each verb. Tables 4.4 and 4.5 illustrate such similarity computation for the selectional equivalents of two senses of *deny* given in (4.2c) and (4.2d). Table 4.4 (left) shows selectors chosen for *confirm* and *contradict*, equivalents for the sense *proclaim*

	refuse	grant	confirm	contradict
refuse	-	0.0983	0.0058	0.0064
grant	-	-	0.0059	0.0000
confirm	-	-	-	0.0487
contradict	-	-	-	-

Table 4.5: Similarity matrix for selectional equivalents of *deny* given in Table 4.4. Similarity values for selectional equivalents of the same sense are underlined. Values are given for top-15 selector lists.

false. Table 4.4 (right) shows selectors chosen for *grant* and *refuse*, equivalents for the sense *refuse to grant*. For comparison, we give conditional probabilities for the same selectors with one of the equivalents of the other sense (*refuse* and *confirm*, respectively).

The resulting similarity scores are shown in Table 4.5. Conditional probability values for the correctly chosen selectors cumulatively insure that the similarity between selectional equivalents of the same sense is higher than their similarity with selectional equivalents of the other sense. This similarity measure thus enables us to differentiate between senses by obtaining clusters of selectional equivalents that can then be used to identify selectors for each of the senses of the target predicate.

4.2.3 Implementation

We used a custom-designed agglomerative clustering engine implemented in C++. The clustering engine allows for easy extension with different scoring schemes, soft/hard clustering implementations, and similarity measures.

Experimentation was conducted with 100M word British National Corpus, using two sets of grammatical relations. The first set was obtained using the Sketch Engine library (Kilgarriff et al., 2004) to which uses a set of regular expressions to extract grammatical relations and index the corpus. The second set was obtained using the Robust Accurate Statistical Parser system (RASP) (Briscoe and Carroll, 2002).

The Sketch Engine parser extracts grammatical relations using regular expressions over pos-tagged text. A number of patterns is defined for each relation, so the `subject` relation, for example, is extracted in both active and passive. Unary, binary, and trinary relations are extracted. Binary relations between headwords and dependents are defined for the argument positions. Trinary relations capture the dependencies with PPs, and link the headword to the head of the NP governed by the dependent prepositional phrase. Unary relations capture particular syntactic configurations, e.g. the fact that a verb is followed by a gerundive. Most of the extracted binary relations, such as `object/object_of`, `subject/subject_of`, `a_modifier/modifies`,

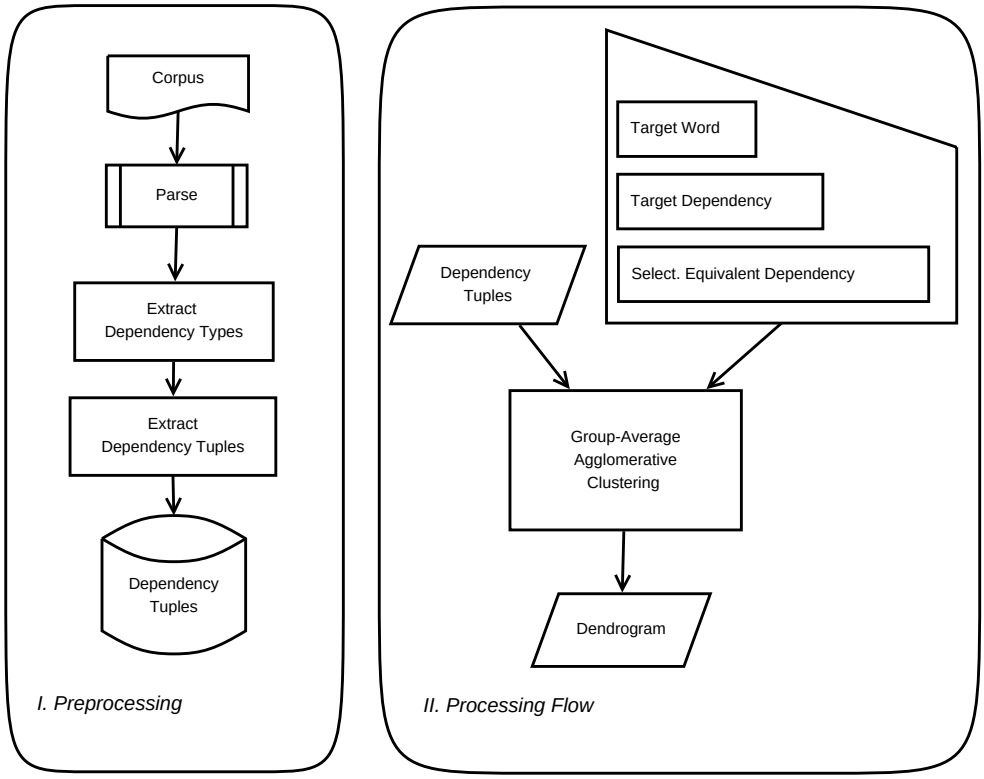

Figure 4.3: Processing Flow

etc. are bidirectional. See Appendix A.2 for a fuller description of the Sketch Engine system.

The set of relations extracted with RASP was compiled as follows:

1. A relation is added to the relation set for each RASP dependency that does not have a third element, i.e., an introducing preposition, conjunction, etc.

   Examples: ncsubj (non-clausal subject), dobj (direct object), etc.

2. For each RASP dependency that occurs with an introducing word, a new relation is added for each introducing word. The introducing word is added to relation name with an underscore.

   Examples: iobj_with (indirect object introduced by "with"), iobj_in (indirect object introduced by "in"), etc.

   Introducing words are lowercased, so iobj_In is equivalent to iobj_in. The "<blank>" symbol used by RASP, when it occurs inside an introducing word, is replaced with underscore: iobj_with<blank>regard<blank>to → iobj_with_regard_to.

3. A relation is added if the introducing word occurs in the first position in the dependency tuple. The only exception is ncsubj_obj (the relation that corresponds to subject of a passive verb).

4. Relation inverse is added for each relation, e.g. iobj_with_inverse is added for iobj_with.

Frequency thresholding was used, so that relations that occur in the corpus less than 5 times (e.g. ncmod_1968) were considered spurious and were not added to the relation set. For more detail about RASP dependencies, see Appendix A.

Once relation set were compiled, populated grammatical relations were extracted from the parsed BNC, with frequency counts for each relation tuple. Processing flow is summarized in Figure 4.3.

4.3 Summary

In this chapter, we have described an unsupervised sense induction system which targets the sense distinctions of the target verb that are linked to the semantics of a particular argument.

The system groups together semantically diverse nouns that activate the same sense of the target verb. This is accomplished by creating contextualized clusters of selectional equivalents for different senses of the target. The results are used to induce a soft clustering for the nouns that occur in the specified argument position. In Chapter 6, we adapt the described system for use in a standard word sense induction task, and test several configurations of the system against the sense-annotated data set described in Chapter 5. Other applications and extensions of the system we presented are discussed in Chapter 7.

Chapter 5

Argument-based Sense Annotation

5.1 Motivation

Semantically annotated corpora are routinely developed for the training and testing of automatic sense detection and induction algorithms. But they do not typically provide a way to distinguish between different kinds of ambiguities. Consequently, it is difficult to perform adequate error analysis for different sense detection systems. Appropriate semantic annotation that would allow one to determine which sense distinctions can be detected better by automatic systems does not need to be highly specific and unnecessarily complex, but requires development of robust generalizations about sense relations.

One obvious conclusion is that data sets need to be explicitly restricted to the instances where humans have no trouble disambiguating between different senses. Thus, prototypical cases can be accounted for reliably, ensuring the clarity of annotated sense distinctions. At face value, imposing such restrictions may appear to negatively influence the usability of the resulting data set in particular applications requiring WSD, such as machine translation or information retrieval. However, this decision impacts most strongly those boundary cases which are not reliably disambiguated by human annotators, and which rather introduce noise into the data set.

In this chapter, we discuss the first attempt at development of the data set that targets specifically one of the main sentence-level features contributing to the disambiguation. Namely, the goal is to target the semantics of the arguments as the source of sense differentiation for a polysemous predicate. We describe how the data set was constructed, and examine the way the annotators dealt with the verbal ambiguities that depend on argument semantics. We further discuss how the relations observed between different senses within a verb's sense inventory influenced annotation decisions.

5.2 Task Description

We were interested specifically in those cases where disambiguation needs to be made without relying on the syntactic frame, and the main source of disambiguation is the semantics of the arguments. Such cases are harder to identify formally in the development of sense inventories and harder for the annotators to disambiguate. For example, phrasal verbs or idiomatic constructions that help identify a particular sense were intentionally excluded from our data set. Thus, for the verb *cut*, one of the senses involves cutting out a shape or a form (e.g. *cut a suit*), but the sentences with the corresponding phrasal form *cut out* were thrown out.

Even so, syntactic clues that contribute to disambiguation in some cases overrule the interpretation suggested by the argument. For example, for the verb *deny*, in *deny the attack*, the direct object strongly suggests a propositional interpretation for *deny* (that the attack didn't happen). However, the use of the ditransitive construction (indicated in the example below by the past participle) overrules this interpretation, and we get the *refuse to grant* sense:[1]

(5.1) Astorre, *denied* his *attack*, had stayed in camp, uneasily brooding.

In fact, during the actual annotation, one of the annotators did not recognize the use of the past participle, and erroneously assigned the *state or maintain something to be untrue* sense to this sentence.

Preparing sense-tagged data for training and evaluation of word sense disambiguation (WSD) systems involves two stages: (1) creating a sense inventory and (2) applying it in annotation. The first stage, which we will refer to as *data set construction* stage involves selecting the set of target words to be annotated, and compiling a sense inventory for each target. The annotation guidelines are then prepared and the data is preprocessed and loaded into the annotation interface. During the *annotation stage*, the target words are disambiguated by the annotators and annotation judgements entered into the database. For our task, the set of targets was comprised by (verb, grammatical relation) pairs.

5.2.1 Data set construction

The data set was developed using the British National Corpus (BNC), which is more balanced than the more commonly annotated Wall Street Journal data. We selected 20 polysemous verbs with sense distinctions that were judged to depend for disambiguation on the semantics of the argument in several argument positions, including

[1] All examples in this chapter are taken from the annotated data set. In some cases, sentence structure was slightly modified for brevity.

direct object (dobj), subject (subj), or indirect object within a prepositional phrase governed by *with* (iobj_with):

dobj: *absorb, acquire, admit, assume, claim, conclude, cut, deny, dictate, drive, edit, enjoy, fire, grasp, know, launch*
subj: *explain, fall, lead*
iobj_with: *meet*

We used the Sketch Engine (Kilgarriff et al., 2004) both to select the verbs and to aid the creation of the sense inventories. The Sketch Engine is a lexicographic tool that lists collocates that co-occur with a given target word in the specified grammatical relation. The collocates are sorted by their association score with the target.

A set of senses was created for each verb using a modification of the CPA technique (Pustejovsky et al., 2004). A set of complements was examined in the Sketch Engine.[2] If a clear division was observed between semantically different groups of collocates in a certain argument position, the verb was selected. For semantically distinct groups of collocates, a separate sense was added to the sense inventory for the target. For example, for the verb *acquire*, a separate sense was added for each of the following sets of direct objects:

(5.2) a. *Take on certain characteristics*
shape, meaning, color, form, dimension, reality, significance, identity, appearance, characteristic, flavor
b. *Purchase or become the owner of property*
land, stock, business, property, wealth, subsidiary, estate, stake

The sense inventory for each verb was cross-checked against several resources, including WordNet, PropBank, Merriam-Webster and Oxford English dictionaries, and existing correspondences in FrameNet (Ruppenhofer et al., 2006; Hiroaki, 2003), OntoNotes (Hovy et al., 2006),[3] and CPA patterns (Hanks and Pustejovsky, 2005; Rumshisky and Pustejovsky, 2006; Pustejovsky et al., 2004). The set of sense inventories for each verb is given in Appendix B.

We performed test annotation on 100 instances, with the sense inventory additionally modified upon examining the results of the annotation. This sense inventory was provided to two annotators, along with 200 sentences for each verb. Each sentence was pre-parsed with RASP (Briscoe and Carroll, 2002), and the head of the target argument phrase was identified. Misparses were manually corrected in post-processing.

[2]See Appendix A.2 for more detail.
[3]Sense inventories released for the 65 verbs made available for SemEval-2007 were used.

5.2.2 Defining the task for the annotators

Data set creation for a WSD task is notoriously hard,[4] as the annotators are frequently forced to perform disambiguation on sentences where no disambiguation can really be performed. This is the case, for example, for overlapping senses, where more than one sense is activated simultaneously (cf. Apresjan, 1973; Pustejovsky and Boguraev, 1993; Rumshisky, 2008). In this task, our goal was to create, for each target word, a set of instances where humans had no trouble disambiguating between different senses.

Two undergraduate linguistics majors served as annotators. The annotators were instructed to mark each sentence with the most fitting sense. The annotators were allowed to mark the sentence as "N/A" and were instructed to do so if (i) the sense inventory was missing the relevant sense, (ii) more than one sense seemed to fit, or (iii) the sense was impossible to determine from the context.[5]

With respect to metaphoric senses, instructions were to throw out cases of creative use where the interpretation was difficult or not immediately clear. The cases where the target grammatical relation was actually absent from the sentence also had to be marked as "N/A" (e.g. for *fire*, sentences without direct object, e.g. *a stolen car was fired upon*). The annotators were also instructed to mark idiomatic expressions and phrasal verbs as "N/A", e.g. for the verb *fall*: *fall from favor, fall through, fall in, fall back, fall silent, fall short, fall in love.*

Disagreements between the annotators were resolved in adjudication by two linguists. The average inter-annotator agreement (ITA) for our data set was computed as a micro-average of the percentage of instances that were annotated with the same sense by both annotators to the total number of instances retained in the data set for each verb. The instances that were marked as "N/A" by one of the annotators (or thrown out during the adjudication) were not included in the computation. The ITA value for our data set was 95%. However, as we will see below, the ITA values do not always reflect the actual accuracy of annotation, due to some common problems with sense inventories.

5.3 Annotation Interface

During the *data set construction* stage, a set of target verbs is selected, and a sense inventory is compiled as described in Section 5.2.1. A set of sentences for each (target verb, grammatical relation) pair was selected randomly from the British National Corpus. Each sentence was automatically parsed with RASP (Briscoe and Carroll,

[4]See, for example, Palmer et al. (2007) for a discussion of this subject.

[5]The full annotation guidelines are given in Appendix C.

2002). During this stage, the sentences was discarded if the target relation was not present.

Annotators were given general annotation instructions (see Appendix B), as well as specific instructions for each verb. Each annotator received a one-hour training session, during which they were asked to read the general annotation instructions, and the sense distinctions for each verb were further explained. The general instructions were also made available at the annotator login page. The annotators were presented with a list of verbs, with a given grammatical relation for each verb, as shown in Figure 5.1.

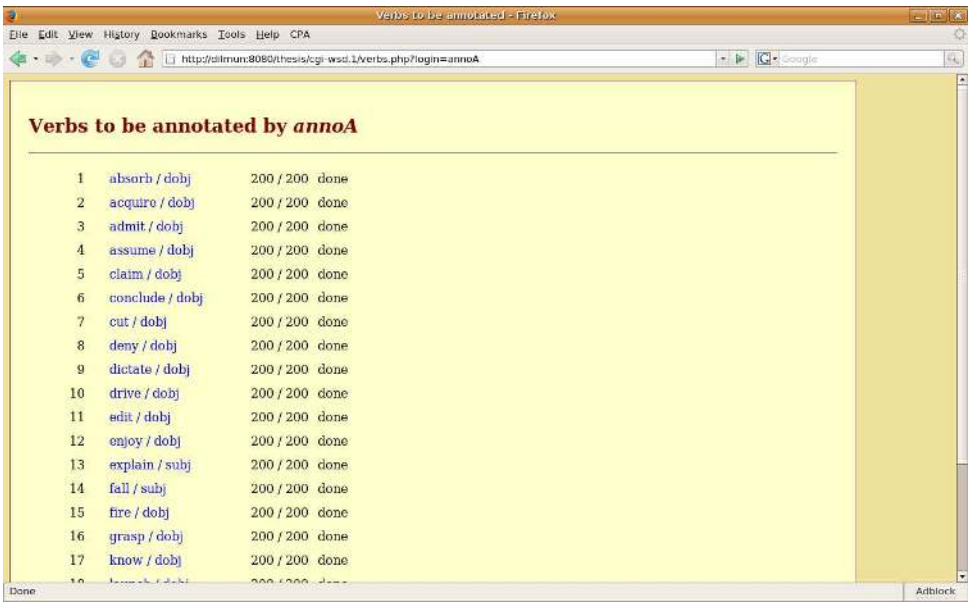

Figure 5.1: Annotation interface: Target selection

The interface then displayed a set of sentences containing the target verb and the chosen grammatical relation. Both the verb and the headword of the dependent noun phrase were highlighted. The annotators were asked to select the most fitting sense of the target verb, or to throw out the example (pick the "N/A" option) if no sense can be chosen either due to insufficient context, because the appropriate sense does not appear in the inventory, or simply no disambiguation can be made in good faith. In further use, the "N/A" category will be split into the following subcategories:

(i) Not enough context determine the sense

(ii) Sense not in inventory

(iii) Borderline case between two available sense

(iv) Metaphoric or creative exploitation of one of the senses

The interface is shown in Figure 5.2. The sense inventories and verb-specific instructions were available at the top of the page for each verb, as in Figure 5.3.

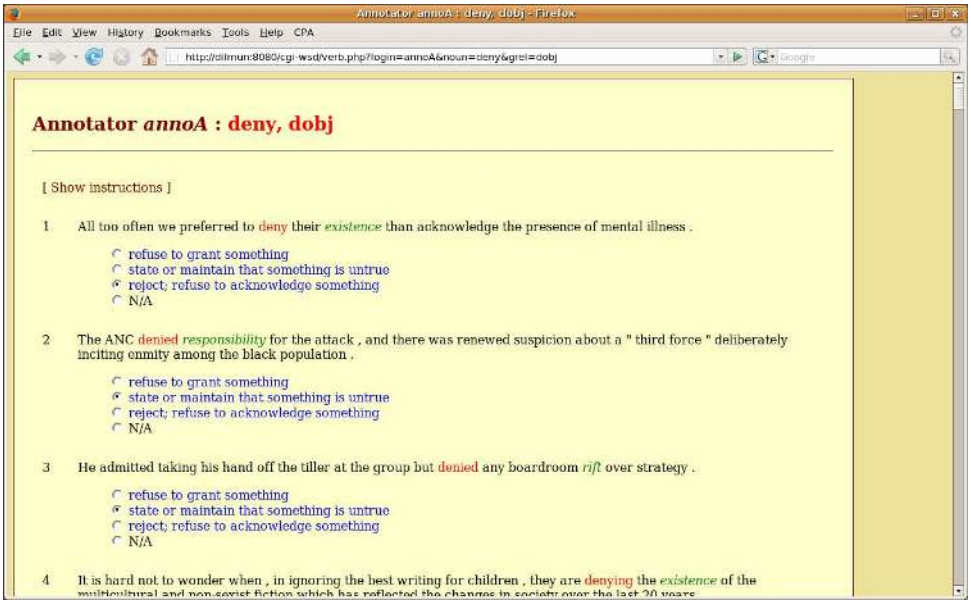

Figure 5.2: Annotation interface: Predicate sense disambiguation for *deny*

After this step is completed by the annotator, the appropriate sense is saved into the database. The senses entered by the annotators for each sentence are displayed and adjudicated in the interface shown in Figure 5.4. The same interface was also used to correct parsing errors, selecting the correct argument where appropriate.

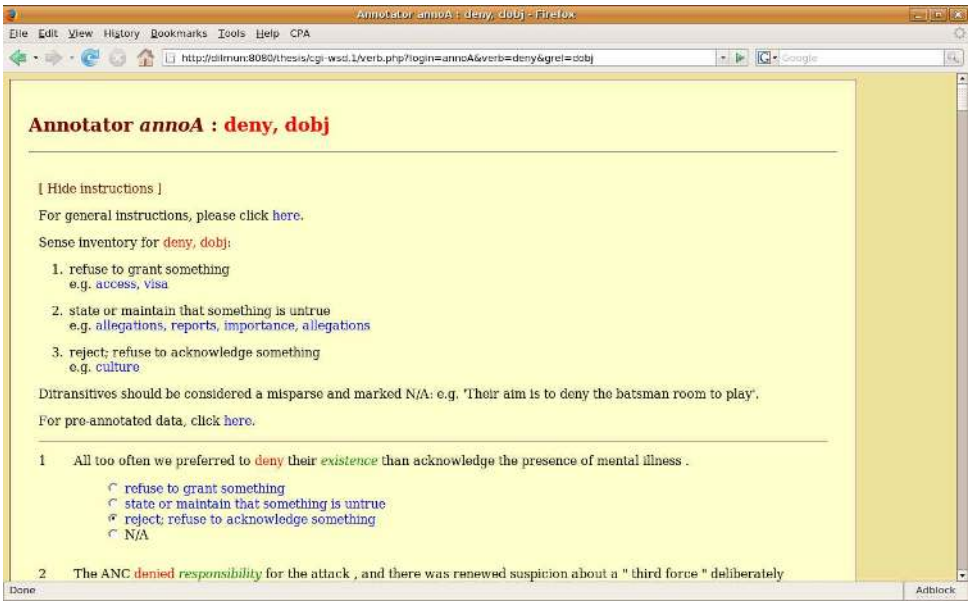

Figure 5.3: Annotation interface: Instructions display

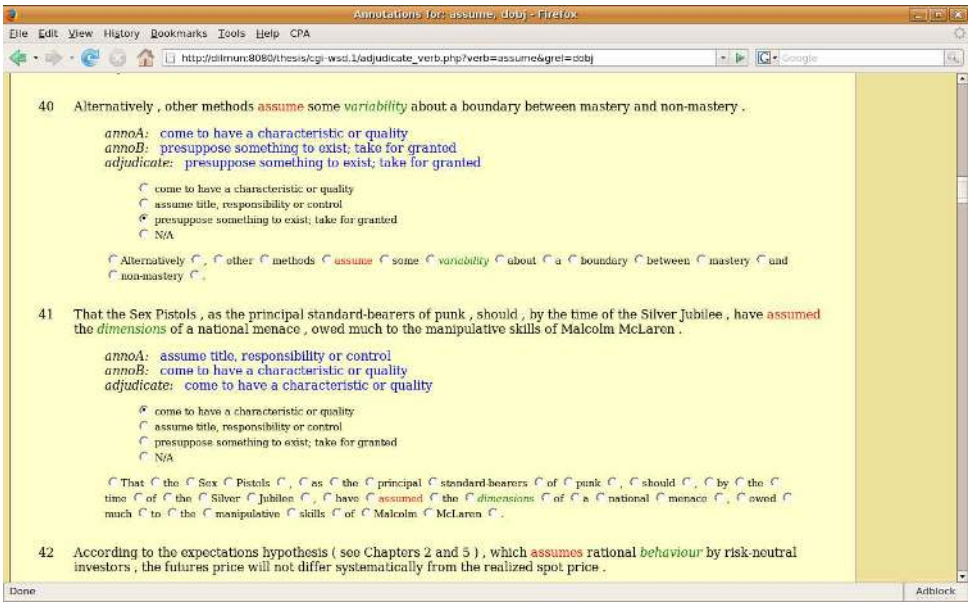

Figure 5.4: Adjudication interface

5.4 Systematic Relations Between Senses

In this section, we discuss the linguistic processes underlying relations between senses within a single sense inventory. We believe that a detailed analysis of these processes should help to account for the annotator's ability to perform disambiguation. Some sense distinctions appear more striking to the annotators, depending on the type of relation involved.

In line with existing approaches to sense relations, we will look at both the linguistic structures involved in sense modification and the productive processes acting on linguistic structures. For the purposes of our present discussion, we interpret the literal (physical, direct) senses to be primary, with respect to more abstract or metaphorical senses.

5.4.1 Argument structure alternations

Some of the most striking differences between the senses are related to the argument structure alternations:

1. Different case roles (frame elements) may be expressed in the same argument position (in this case, direct object), corresponding to different perspectives on the same event. For example, direct object position of the verb *drive* may be filled by VEHICLE, DISTANCE, or PHYSOBJ giving rise to three distinct senses: (i) *operate a vehicle controlling its motion*, (ii) *travel in a vehicle a certain distance*, and (iii) *transport something or someone*. Similarly, for the verb *fire*, PROJECTILE or WEAPON in direct object position give rise to two related senses: (i) *shoot, discharge a weapon*, (ii) *shoot, propel a projectile*.

2. The distinction between propositional and non-propositional complements, as for the verbs *admit* and *deny* in (5.3) and (5.4):

(5.3) a. *admit defeat, inconsistency, offense*
(*acknowledge the truth or reality of*)
b. *admit patients, students*
(*grant entry or allow into a community*)

(5.4) a. *deny reports, importance, allegations*
(*state or maintain to be untrue*)
b. *deny visa, access*
(*refuse to grant*)

3. There is a mutual dependency between subcategorization features of the complements in different argument positions. For example, the [+animate] subject may

combine with specific complements not available for [−animate], as for the two senses of *acquire*: (i) *learn* and (ii) *take on certain characteristics*. Compare NP_{subj} [-animate] *acquire* NP_{dobj} (*language, manners, knowledge, skill*) vs. NP_{subj} [−animate] *acquire* NP_{dobj} (*importance, significance*). Similarly, for *absorb*, compare NP_{subj} [±animate] *absorb* NP_{dobj} (*substance*) and NP_{subj} [+animate] *absorb* NP_{dobj} (*skill, information*). Note that, as one would expect, such dependencies are inevitable even despite the fact that our data set was developed specifically to target sense distinctions dependent on a single argument position.

5.4.2 Event structure modification

Event structure modifications (i.e. operations affecting aspectual properties of the predicate) are another source of sense differentiation. Two cases appear most prominent:

1. The event structure is modified along with the characteristics of the arguments. For example, for *enjoy*, compare *enjoy skiing, vacation* (DYNAMIC EVENT) with *enjoying a status* (STATE). Similarly, for *lead*, compare *a person leads smb somewhere* (PROCESS) vs. *a road* (PATH) *leads somewhere* (STATE); for *explain*, compare *something or somebody explains smth* (= *clarifies, describes, makes comprehensible*, PROCESS) vs. *something* [−inanimate, +abstract] *explains something* (= *is a reason for something*, STATE); for *fall*, compare PHYSOBJ *falls* (TRANSITION or ACCOMPLISHMENT) vs. *a case falls into a certain category* (STATE).

2. The aspectual nature of the predicate is the only semantically relevant feature that remains unchanged after consecutive sense modifications. For example, the ingressive meaning of 'beginning something' is preserved in shifting from the physical sense of the verb *launch* in *launch a missile* to *launch a campaign* and *launch a product*.

5.4.3 Lexical semantic features

Sense distinctions often involve deeper semantic characteristics of the verbs which could be accounted for by means of lexical semantic features such as qualia structure roles in Generative Lexicon:[6]

1. Consider how the meaning component 'manner of motion' (typically associated with the agentive role) gets transformed in the different senses of *drive*. It is obviously present in the physical uses of *drive* (such as *operate a vehicle*, *transport something or somebody*, etc.), but is completely lost in *motivate the progress of* (as in *drive the economy, drive the market forward*, etc.). The value of the agentive role of *drive*

[6]We will use the terminology from Generative Lexicon (Pustejovsky, 1995; Pustejovsky, 2007) to discuss lexical semantic properties, such as *qualia roles*, *complex* and *functional types*, and so on.

becomes underspecified or semantically weak, so that the overall meaning of *drive* is transformed to *cause something to move*.

2. Information about semantic type contained in qualia structure allows apparently diverse elements to activate the same sense of the verb. For instance, the verb *absorb* in the sense *learn or incorporate skill or information* occurs with direct objects such as *values, atmosphere, information, idea, words, lesson, attitudes, culture*. The requisite semantic component is realized differently for each of these words. Some of them are complex types with INFORMATION as one of the constituent types: *words* (ACOUSTIC/VISUAL ENTITY • INFO), *lesson* (EVENT • INFO). Others, such as *idea*, are polysemous, with one of the senses being INFORMATION. Cases like *culture* and *values* are more difficult, but since they refer to knowledge, the INFORMATION component is clearly present. Consequently, the annotators are able to identify the corresponding sense of *absorb* with a high degree of agreement.

5.4.4 Metaphor and metonymy

Meaning transformations in our corpus often involve metaphor and metonymy. Below are some of the conventionalized extensions with a metaphorical flavor:

(5.5) a. *grasp object* vs. *grasp meaning*
b. *launch object* vs. *launch an event (campaign, assault)* or *launch a product (newspaper, collection)*
c. *meet with a person* vs. *meet with success, resistance*
d. *lead somebody somewhere* vs. *lead to a consequence*

Note that the metaphorical extensions in (5.5) involve abstract or continuous objects (*meaning, assault, success, consequence*), which in turn cause event structure modifications (*lead* as a process vs. *lead* as a state). Thus, the processes and structures we are dealing with are clearly interrelated.

The metonymical process can be exemplified by two senses of *edit*: *make changes to the text* and *supervise publication*, which are in a clear contiguity relationship.

One of the effects of the metaphorization and progressive emptying of the primary (physical, concrete) senses is the distinction between generic and specific senses. For example, compare *acquire land, business* (specific sense) to *acquire an infection, a boyfriend, a following*, which refers to some extremely light generic association. Similar process is observed for the semantically weak sense of *fall*, *be associated with or get assigned to a person or location or for event to fall onto a time*:

(5.6) Birthdays, lunches, celebrations *fall* on a certain date or time
Stress or emphasis *fall* on a given topic or a syllable

Responsibility, luck, suspicion *fall* on or to a person

The specificity often involves specialization within a certain domain:

(5.7) a. *conclude* as *finish* vs. *conclude* as *reach an agreement* (Law, Politics)
b. *fire* as *shoot a weapon or a projectile* vs. *fire* as *kick or pass an object of play in sports* (Sport)

Thus, when concluding a *pact* or an *agreement*, a certain EVENT is also being finished (negotiation of that agreement), necessarily with a positive outcome.

In the following section, we will show how different kinds of relations between senses influence disambiguation carried out by the annotators. In particular, we look at different sources of disagreement and annotator error as determined in adjudication.

5.5 Analysis of Annotation Decisions

As we have seen above, in many cases disambiguation is impossible due to the nature of compositionality. Also, as there are no clear answers to a number of questions concerning sense identification, the annotators deal with sense inventories that are imperfect. Results of the disambiguation task carried out by the annotators reflect all these defects.

In cases when a specific meaning from the data set is not included into the sense inventory (e.g. due to its low frequency or extreme fine-grainedness) the annotators may use a more general meaning or pick the closest meaning available. For example, within the sense inventory for *fire*, there was no separate gloss for *fire an engine*. Annotator A in our experiment chose the closest specific meaning available, and Annotator B marked it with a more generic sense:

(5.8) Engineers successfully *fired* thrusters to boost the research satellite to an altitude of 507 km.
annoA: *shoot, propel a projectile*
annoB: *apply fire to*

As mentioned in Section 3.2.5, even when the appropriate specific sense is available, annotators frequently chose the more generic sense in its place, as in (5.9)–(5.11), and also in (3.20).

(5.9) Several *referrals fell* into this *category*.
annoA: *be associated with or get assigned to a person or location or for event*

to fall onto a time
annoB: *be categorized as or fall into a range*

(5.10) The terrible *silence* had *fallen*.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *for a state (such as darkness or silence) to come, to commence*

(5.11) He *acquired* a *taste* for performing in public.
annoA: *become associated with something, often newly brought into being*
annoB: *become associated with something, ...*
correct: *learn*

Note that in the example we gave in (3.20) this decision was probably motivated by the annotators' uncertainty about the semantic ascription of the relevant argument (*coastline* is not a prototypical owned property). The generic sense seems to be the safest option to take for the annotators, as compared to taking a chance with a specific meaning. Due to its low degree of semantic specification, the generic sense is potentially able to embrace almost every possible use. This is not a desirable outcome because the generic senses are introduced in the inventory to account only for semantically underspecified cases. For instance, *become associated with something, often newly brought into being* is appropriate for *acquire a grandchild*, but not for *acquire a taste* or *acquire a proficiency*.

Remarkable variation is also observed with respect to **non-literal uses** as discussed in Section 5.4.4. For example, in (5.12) and (5.13) abstract NPs *panic* and *imbalance of forces* are equated with *energy or impact* by one annotator and with *substance* by the other.

(5.12) Her *panic* was *absorbed* by his warmth.
annoA: *absorb energy or impact*
annoB: *absorb substance*

(5.13) Alternatively, *imbalance* of forces can be *absorbed* into the body.
annoA: *absorb energy or impact*
annoB: *absorb substance*

In some cases, the literal and the metaphoric senses are activated simultaneously resulting in ambiguity (cf. Cruse (2000)):

(5.14) For over 300 years this waterfall has provided the energy to *drive* the *wheels* of industry.
annoA: *motivate the progress of*
annoB: *provide power for or physically move a mechanism*

(5.15) But fashion changed and the short *skirt fell* – literally – from favour and started skimming the ankles.
annoA: *lose power or suffer a defeat*
annoB: N/A

(5.16) She was delighted when the *story* of Hank *fell* into her lap.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *physically drop; move or extend downward*

The impact of **subcategorization features** on disambiguation (cf. Section 5.4.1 para 3) is illustrated in (5.17).

(5.17) The reggae tourist can easily *absorb* the current reggae *vibe*.
annoA: *absorb energy or impact*
annoB: *learn or incorporate skill or information*

Both interpretations chosen here (*absorb energy or impact* and *learn or incorporate skill or information*) were possible due to the animacy of the subject, which activates two different subcategorization frames and subsequently two different senses.

Typically, cases where **semantic type** of the relevant arguments (cf. Section 5.4.3 para 2) is not clear result in annotator disagreement:

(5.18) The AAA *launched* education *programs*.
annoA: *begin or initiate an endeavor* (EVENT)
annoB: *begin to produce or distribute; start a company* (PRODUCT)

(5.19) France plans to *launch* a remote-sensing *vehicle* called Spot.
annoA: *physically propel into the air, water or space* (PHYSOBJ)
annoB: *begin to produce or distribute; start a company* (PRODUCT)

The two cases above are interesting in that both *program* and *vehicle* are ambiguous and can be analyzed semantically as members of different semantic classes.

This is what the annotators in fact do, and as a result, ascribe them to different senses. *Program* can be categorized as EVENT ('series of steps') or as INTELLECTUAL ACTIVITY PRODUCT ('document or system of projects'). It is a complex type, i.e. it is an inherently polysemous word that represents at least two different semantic types. *Vehicle*, in turn, is a functional type: on the one hand, it represents an entity with certain formal properties (PHYSOBJ interpretation), on the other hand, it is an artifact, with a prominent practical purpose (PRODUCT interpretation).

In fact, most problems the annotators had with the task are due to the inherent semantic complexity of words such as *vehicle* and *program* in (5.18) and (5.19) and to the existence of boundary cases, where the relevant noun does not properly belong to one or another semantic category. This is the case with *panic*, *imbalance* or *reggae vibe* in (5.12), (5.13), and (5.17), and also with *taste* and *coastline* in (5.11) and (5.2).

In some of these cases, other contextual clues may come into play and tip the balance in favor of one or another sense. Note that disambiguation was influenced by a **wider context** even despite the intentionally restrictive task design (targeting a particular syntactic relation for each verb). For instance, in (5.20), **domain-specific clues** referring to war or military conflict (such as *rebel control*) could have motivated Annotator B's decision to ascribe it to the sense *lose power or suffer a defeat* (even though a road is not typically an entity that can lose power), while the other annotator chose a more generic meaning:

(5.20) The *road fell* into rebel control.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *lose power or suffer a defeat*

Other pragmatic and discourse-oriented clues played a role, in particular, positive and negative connotation of the senses and the relevant arguments, as well as the temporal organization of discourse. For example, in (5.21) and (5.22), positive or neutral interpretation of *wave of immigrants* and *change* could have led to the choice of *take in or assimilate* and *learn or incorporate skill or information* senses, while the negatively-colored interpretation might explain the choice of the *bear the cost of* sense.

(5.21) ..help *absorb* the latest *wave of immigrants*.
annoA: *bear the cost of; take on an expense*
annoB: *take in or assimilate, making part of a whole or a group*

(5.22) For senior management an important lesson was the trade unions' capacity to *absorb change* and to become its agents.

annoA: *learn or incorporate skill or information*
annoB: *bear the cost of; take on an expense*

Temporal organization of a broader discourse is another important factor. For example, for the verb *claim*, the senses *claim the truth of* and *claim property you are entitled to* have different presuppositions with respect to preexistence of the thing claimed. In (5.22), due to the absence of a broader context, the annotators chose two different temporal reference interpretations. For Annotator B, *success* was something that has happened already, while for A this was not clear (*success* might have been achieved or not):

(5.23) One area where the government can *claim* some *success* involves debt repayment.
annoA: *come in possession of or claim property you are entitled to*
annoB: *claim the truth of*

5.6 Summary

In this chapter, we have described the construction of a data set targeting the semantics of a particular argument as the source of sense differentiation for the predicate. We have given an overview of different types of sense relations in polysemous predicates and analyzed their effect on different aspects of the annotation task, including sense inventory design and execution of the WSD annotation.

In the next chapter, we use a subset of the resulting data set to test the clustering algorithm described in Chapter 4. The imperfections inherent in the production of such annotated data are quite clear, but as we show in the next chapter, the resulting set compares favorably with state-of-the-art sense-tagged data such as the data used in the recent Semeval competition (Agirre et al., 2007). Further characteristics of the data set, as used in testing as a standard sense-tagged corpus, are discussed in Section 6.3.

Chapter 6

Evaluation via Word Sense Induction

6.1 Motivation

In Chapter 4, we described a fully unsupervised sense induction system that analyzes how different arguments contribute to disambiguation, looking at one argument position at a time.

The constructed system has a number of advantages. To get an idea of how well it performs, we evaluated it in a standard sense induction setting. Such evaluation requires a manually constructed resource to evaluate the induced groupings. We use a subset of the sense-tagged data set described in Chapter 5. The subset consisted of 15 polysemous verbs selected for having a higher inter-annotator agreement.

As mentioned above, using the standard sense-tagged data sets in evaluation is not feasible in this case because most existing sense-annotated corpora tend to conflate different kinds of contextual information. Initially, we considered using the data set used in the last SEMEVAL competition for the WSD and WSI tasks (Tasks 17 and 2, respectively) (Agirre et al., 2007). This option was especially attractive, since we could then compare the performance of our system directly to the performance of the state-of-the-art sense induction systems that participated in Task 2. However, the average per-verb entropy of the SEMEVAL data set is 0.92, suggesting that the most frequent sense dominated the data set for many of the chosen verbs. In fact, out of the 65 verbs used in the WSI task, 11 verbs had only one sense in the combined test and training data set. Such distribution across the senses is problematic, especially since the evaluation schemes used in Task 2 relied to a large extent on the most frequent sense to assess the systems' performance.[1]

For the above reasons, we do not use the SEMEVAL Task 2 data set for eval-

[1]For further discussion of this, see Section 6.4.

uation directly. Instead, we perform an indirect comparison using the data set we developed. We use several measures to evaluate the clustering solution quality, including two measures we introduce here, and compare our system's performance to the performance of the SEMEVAL Task 2 systems. We report the results for the following four configurations: MI-FACT-PROD, MI-FACT-PROD-PROD, MI-PROD, and MI-PROD-PROD. Relative to the baselines, our system outperforms the best system in the SEMEVAL Task 2 on two out of three measures.

The rest of this chapter is organized as follows. Section 6.2 describes the way the system is adapted to use in a standard WSI task. Section 6.3 gives detailed characteristics of the data set used for for evaluation. In Section 6.4, we describe the evaluation schemes we chose and discuss results.

6.2 WSI Algorithm

The data set we used for evaluation in this chapter consisted of verbs, and the direct object relation was used both for the target and for its selectional equivalents. All the computations were performed over the 100M word British National Corpus (BNC, 2000). We used RASP (Briscoe and Carroll, 2002) to extract grammatical relations.

For each target word t and relation R, we follow the steps described in Chapter 4, Section 4.2.1, building a dendrogram for each of the configurations listed in Table 4.2. We then rank each cluster, and use the high-ranking clusters in the WSI task as described below.

6.2.1 Cluster rank

We sort all the nodes in the dendrogram by computing the following score for each node C_i:

$$rank(C_i) = IntraAPS(C_i) \cdot \log(|C_i|) \cdot \log(\sum_{s \in C_i} f_i(s)) \tag{6.1}$$

where $f_i(s)$ is the score assigned to the selector within cluster C_i, $|C_i|$ is the number of elements in C_i, and $IntraAPS(C_i)$ is the average pairwise similarity between the elements of the cluster.

In the present experiments, we used the top 20 clusters that maximized this score.

6.2.2 Selector-cluster association

Using the obtained clusters, we can estimate which sense of the target a selector is likely to occur with. We compute an association score for each of the chosen clusters C_i and selector s:

$$assoc(s, C_i) = \frac{\sum_{w \in C_i} mi(s, Rw)}{|C_i|} \tag{6.2}$$

where $mi(s, Rw) = \log \frac{P(s,R,w)}{P(s)P(R,w)}$.

The resulting score indicates how likely selector s is to pick the sense of the target associated with C_i. The difference between the scores obtained for different senses with a given selector indicates how strongly that selector tends to prefer one of the senses. If the difference is small, the selector either must equally likely select for either of the senses, or select for both senses at once.

6.2.3 Using clusters in a WSI task

The obtained dendrogram was adapted for use in the standard word sense induction task as follows. Given a set of sentences containing the target word, we extracted the selector for the appropriate grammatical relation. For each selector, we then computed the selector-cluster association score with each of the high-ranking clusters. The sentences containing selector s are tagged with the cluster that has maximum $assoc(s, C_i)$. The sentences that are tagged with intersecting clusters (i.e. clusters containing at least some of the same selectional equivalents of the target) are then grouped together.

This method has an obvious disadvantage relative to the full WSI systems, namely, that we do disambiguation based on only one selector. Consequently, we would expect it to do poorly in situations where a larger context is required for disambiguation.

Here are the clusters obtained for the verbs *conclude* and *grasp* using this method:

verb: conclude
gloss #1: finish
cluster: *begin-v continue-v resume-v prolong-v start-v commence-v open-v initiate-v reopen-v re-open-v*
selectors: negotiation-n, discussion-n, investigation-n, proceedings-n, conversation-n, inquiry-n, talk-n, debate-n, friendship-n, deliberation-n, exploration-n, round-n, argument-n, conquest-n, tour-n, ...
gloss #2: reach an agreement
cluster: *sign-v renegotiate-v agree-v negotiate-v*
selectors: deal-n, pact-n, contract-n, treaty-n, agreement-n, covenant-n, settlement-n, ceasefire-n, arrangement-n, armistice-n, truce-n, ...

verb: grasp
gloss #1: understand, comprehend
cluster: *appreciate-v recognise-v recognize-v realise-v realize-v assess-v demonstrate-v reflect-v illustrate-v explain-v understand-v acknowledge-v underline-v emphasize-v stress-v emphasise-v*
selectors: importance-n, nature-n, significance-n, potential-n, value-n, difference-n, extent-n, fact-n, point-n, complexity-n, implication-n, relationship-n, principle-n, effect-n, meaning-n, situation-n, truth-n, reality-n, concept-n, role-n, aspect-n, necessity-n, idea-n, ...

gloss #2: grab hold of something
cluster: *put-v hold-v thrust-v touch-v raise-v rest-v lift-v rub-v*
selectors: hand-n, arm-n, chin-n, elbow-n, finger-n, shoulder-n head-n, leg-n, receiver-n, knife-n, wrist-n, hair-n, back-n, sword-n, ...

6.3 Data Set

We tested our system on the subset of the data set described in Chapter 5 We selected 15 polysemous verbs with sense distinctions that were judged to be robust in adjudication. The following verbs were selected for sense distinctions linked with the direct object position:

dobj: *absorb, acquire, admit, assume, conclude, cut, deny, dictate, drive, edit, enjoy, fire, grasp, know, launch*

The inter-annotator agreement (ITA) for this set was 95%, computed as a micro-average as described in Chapter 5, Section 5.2.2. Table 6.1 shows the following characteristics for each verb: 1) ITA (percentage of instances where the annotators selected the same sense for the verb), 2) MFS (percentage of instances that belong to the most frequent sense), 3) the number of senses and number of instances, and 4) entropy of the distribution of instances across senses. The last row of each column gives the average for the column, weighted by the number of instances for each verb.[2]

To determine how well the verbs in our data set could be disambiguated by a supervised system relying solely on nouns in direct object position, we also ran on our data a Maximum Entropy classifier with 10-fold cross-validation.[3] The obtained accuracy values are shown in Table 6.1.

6.4 Evaluation

Our system uses semantics of a single argument to do the disambiguation, but since the verbs in our data have been selected for effectiveness of single-argument semantics disambiguation, it is reasonable to compare the performance of our system to that of the general sense induction systems. One handicap that such evaluation imposes on our system is that since no other context is available, one selector can only be associated with one sense of the target. If that selector can activate more than one sense, i.e. if it is associated with more than one cluster of selectional equivalents, our

[2]Average number of annotated instances in our data was similar to the OntoNotes data set used in the 2007 SEMEVAL competition (Agirre et al., 2007).

[3]We used the Maximum Entropy classifier from the CARAFE project available at `http://sourceforge.net/projects/carafe`.

Word	No. Senses	No. Inst.	ITA %	Entropy	MFS	MaxEnt accuracy	F-measure		
							random	1c1word	mi-fact-prod
absorb	7	196	92.4	2.49	.30	.58	.20	.33	.36
acquire	4	186	92.1	1.86	.44	.44	.30	.45	.59
admit	2	163	98.7	1.00	.53	.71	.51	.67	.74
assume	3	191	90.8	1.55	.45	.73	.39	.52	.48
conclude	2	178	97.5	0.96	.62	.89	.55	.68	.51
cut	4	166	92.3	1.33	.58	.51	.49	.61	.78
deny	3	190	97.2	1.49	.49	.62	.38	.54	.55
dictate	2	193	98.9	0.53	.88	.97	.79	.85	.62
drive	11	174	97.6	2.64	.41	.40	.23	.34	.39
edit	2	176	98.0	0.98	.57	.82	.57	.67	.62
enjoy	2	193	86.2	0.93	.66	.70	.57	.70	.53
fire	6	162	97.3	1.87	.54	.73	.37	.49	.58
grasp	3	178	97.6	1.25	.49	.84	.45	.61	.85
know	2	172	92.6	0.98	.58	.79	.54	.67	.56
launch	3	196	89.9	1.24	.63	.74	.52	.62	.66
Average	3.73	180.9	94.5	1.41	.545	.699	.457	.584	.586

Table 6.1: Per-word characteristics of the data set and system performance

only option is to choose the cluster with which it has the strongest association. We found that our system performed well even despite this handicap.

In Task-2 of SEMEVAL-2007, the participant sense induction systems were evaluated using Wall Street Journal data annotated with OntoNotes senses (Hovy et al., 2006). While we could not re-use that data set with our system, we performed a set of comparisons of our system's performance relative to the characteristics of our data set.

SEMEVAL Task-2 used two kinds of evaluation: supervised and unsupervised. Supervised evaluation divided the data into training and test, and mapped each cluster to the sense that was dominant for the elements of that cluster. Effectively, each cluster was associated with the sense that maximized that cluster's precision. Multiple clusters were thus allowed to be mapped to the same sense. The resulting mapping was applied to compute accuracy on the test set. Under such evaluation, the obtained accuracy depends strongly on the majority baseline for each word in the data set: there is no penalty for splitting a dominant sense into several clusters. For this reason, we chose not to use this evaluation method.

In the unsupervised evaluation, Van Rijsbergen's F-measure was used to rank the participating systems. This method used a set-matching evaluation technique optimizing F-measure. The set matching stage found the optimum cluster for each sense, and averaged the F-measure of the best-matching cluster across all senses. Two

relevant baselines were computed for the data set: (1) all instances for the given target word clustered together (*1cluster1word*) and (2) each instance treated as a separate cluster (*1cluster1inst*). Under this metric, the *1cluster1word* baseline (all occurrences of the target word grouped together) outperforms all the clustering systems that competed in the task. This is due to the known problems with this measure (Meila, 2003).

A number of other metrics are available, and we have reviewed some of them earlier (cf. Ch. 2). We were interested in metrics that would support certain reasonable constraints, such as giving a lower score to the solution that merges two clusters that correspond to different senses, or unnecessarily splits a single sense. We also wanted to see the comparison produced by metrics that (1) do not require the set matching to evaluate a particular clustering solution, and/or (2) consider the quality of mapping in both directions.

We used the following metrics to evaluate the performance of our system: (1) F-measure as used in SEMEVAL sense induction task (2) BCubed P&R (Amigó et al., 2008) (3) mutual information as used in Meila (2003). We review the latter two measures below:

"BCubed" measures: We used the harmonic mean of BCubed precision and recall, which are defined for a given clustering solution C and a sense assignment solution S on data set D as follows:

$$\textbf{BCubed Precision} = \frac{\sum_e \frac{|C(e) \cap S(e)|}{|C(e)|}}{n}$$

$$\textbf{BCubed Recall} = \frac{\sum_e \frac{|C(e) \cap S(e)|}{|S(e)|}}{n}$$

where $e \in D$ is an element of the data set, $C(e)$ is the cluster to which e belongs, and $S(e)$ is the sense category to which e belongs, and $n = |D|$.

Entropy/MI measures: We used the standard mutual information measure of two variables defined by the clustering solution and the sense assignment $I(C, S)$ in the way delineated in Meila (2003):

$$I(C, S) = \sum_{i,j} P(i, j) \log \frac{P(i, j)}{P(i)P(j)}$$

where $c_i \in C$ is a cluster from the clustering solution C, and $s_j \in S$ is a sense from the sense assignment S, and $P(i, j) = \frac{|c_i \cap s_j|}{n}$. Recall from Ch. 2 that the range for $I(C, S)$ depends on the entropy values of the two variables, $H(C)$ and $H(S)$:

$$0 \leq I(C,S) \leq \min(H(C), H(S))$$

Since we needed to perform comparisons across different data sets, we used $I(C,S)$ normalized by $\max(H(C), H(S))$:

$$\textbf{NormalizedMI} = \frac{I(C,S)}{\max(H(S), H(C))}$$

which allowed us to retain the $(0,1)$ range and certain other desirable properties, such as:

NormalizedMI(*1c1word*,S)= 0
NormalizedMI(*1c1inst*,S)= $H(S)/\log n$
NormalizedMI(S,S)= 1

We computed the F-measure based metric, the BCubed and NormalizedMI metrics both for our system and for the SEMEVAL data.[4] Since our data set included only verbs, we recomputed the metrics separately for the verbs in the SEMEVAL data set, based on the published clustering solutions for each participating system.

Table 6.2 summarizes the values obtained for these metrics by four configurations of our system that use the product of two association scores as the combiner function $\psi(assoc_R(s,w), assoc_R(s,t))$. Table 6.3 gives the values obtained for the same metrics for each of the systems in SEMEVAL Task-2. The reported values in both tables are averages across all target words in the data set. To aid comparison across data sets, next to the actual value obtained by each system, we give the ratio of that value to the best performing baseline.

The verbs in our test data set have a significantly higher degree of polysemy compared to the SEMEVAL data. While the average number of senses per verb in our data and in SEMEVAL data is very similar (3.73 and 3.54, respectively), the distributions of senses differ. The average per-verb entropy for our data set is 1.4, while for SEMEVAL data it is 0.9. Consequently, our data has much lower majority baseline and is potentially more difficult to classify. Note that the average number of instances per target in our data set was similar to the SEMEVAL data set, so the higher value of *1c1word* baseline for NormalizedMI reflects only the difference in the entropy of the annotated data.

Table 6.1 shows the F-measure values obtained for two baselines and for our best-performing configuration (*mi-fact-prod*), for each verb in our data set. The random baseline was computed in the following way: for each verb, we randomly

[4]We use the standard entropy definition (Cover and Thomas, 1991), so unlike in the definition used in SEMEVAL Task-2 (Zhao and Karypis, 2004), the terms are not multiplied by the inverse of the log of the number of senses.

Variant	F-measure		BCubed		Norm. MI	
		% 1c1w		% 1c1w		% 1c1i
1c1inst	.038	6.5	.040	6.7	.188	100
1c1word	.584	100	.599	100	0	0
mi-fact-prod	.586	100.3	.522	87.1	.138	73.4
mi-fact-prod-prod	.572	97.9	.540	90.2	.061	32.4
mi-prod	.504	86.3	.439	73.3	.103	54.8
mi-prod-prod	.544	93.2	.469	78.3	.101	53.7

Table 6.2: Performance of our system for different clustering configurations

System	F-measure		BCubed		Norm. MI	
		% 1c1w		% 1c1w		% 1c1i
1c1inst	.035	4.6	.039	5.0	.118	100
1c1word	.755	100	.776	100	0	0
I2R	.528	69.9	.505	65.1	.051	43.2
UBC-AS	.750	99.3	.769	99.1	.005	4.2
UMND2	.640	84.8	.638	82.2	.006	5.1
UOY	.383	50.7	.253	32.6	.048	40.7
upv_si	.607	80.4	.520	67.0	.044	37.3

Table 6.3: SEMEVAL Task-2 system performance

split the instances into clusters of the same number and size as the sense classes in the annotated data, and calculated the resulting F-measure, averaged over 10 runs.

6.5 Summary

In this chapter, we described how our system can be adapted for use in a standard WSI setting. We evaluated our system on the data set described in Ch. 5. We used three different measures for assessing the quality of the clustering solution: (1) F-measure (Zhao and Karypis, 2004), (2) the harmonic mean of **BCubed Precision and Recall** (Amigó et al., 2008), and (3) **NormalizedMI**, a measure we proposed based on the definition of mutual information for a clustering solution and a sense assignment (Meila, 2003). We computed these measures for the verbs in SEMEVAL data set, based on the published clustering solutions for each system, and performed a comparison with our system, relative to the baselines. As we can see in Tables 6.3 and 6.2, our best configuration outperforms the best SEMEVAL system on both F-measure and NormalizedMI, despite the fact that our system can only associate one sense with all instances of a given selector.

Note that under the evaluation scheme we used, selectional properties of the target verb were analyzed by using selectional equivalents with the same grammatical relation. That is, if the target verb was disambiguated based on the semantics of the direct object, the representation for its selectional equivalents was also computed using direct objects. As we mentioned previously, this does not need to be the case, since our system can model selectional preferences in a given argument position using any other grammatical relation (e.g., subjects or indirect object can be modeled with direct objects, and so on).

Chapter 7

Computational and Theoretical Extensions

The system for clustering selectors of polysemous words which we presented in Chapter 4 is clearly not limited to dealing with selectional properties of polysemous verbs. The same principles apply to other cases of regular polysemy that are difficult to resolve with the methods using non-contextualized distributional similarity. The information about selectional equivalency, as captured by the derived bipartite cluster trees, can clearly be useful in a number of applications. In this chapter, we discuss the way it can be applied dot objects (Pustejovsky, 1995) and to the disambiguation of NPs with semantically weak head nouns.

7.1 Sense Selection in Dot Nominals

In the Generative Lexicon (GL) (Pustejovsky, 1995) knowledge representation framework, complex types (dot objects) are introduced to account for certain types of inherent polysemy. In this section, we discuss some aspects of selectional behavior of dot objects in corpus and then illustrate how bipartite contextualized clustering can be used to identify selector contexts specific to the component types of the dot.

We begin by examining the relevant data. We then illustrate how selector contexts for dot nominals can be clustered according to the selected type.

7.1.1 Data Analysis for Dot Objects

Complex types are introduced in GL as a mechanism for dealing with selectional behavior of nouns such as *lunch* (EVENT • FOOD) and *newspaper* ((PHYS • INFO) • ORGANIZATION). The contexts in which complex types occur may select for any of the simple types that make up the complex type.

(7.1) a. I have my *lunch* in the backpack. (FOOD)
b. Your *lunch* today was longer than usual. (EVENT)

For a dot nominal, the senses that correspond to the simple types are connected in a regular and well-defined manner. Some examples of complex types are given in Table 7.1.[1] Complex types typically allow *multiple selection*:

(7.2) We had a *delicious* (FOOD) *leisurely* (EVENT) *lunch*.

There also exist contexts that select specifically for the complex type of each kind. Thus, for some of the complex types there also seem to exist gating predicates (Pustejovsky, 2007) whose selectional specification may specify a transition between two simple types that make up the complex type. For example, food preparation predicates (e.g. *poach, steam, braise, cook*) are gating predicates for such complex types as ANIMAL • FOOD:

(7.3) She wouldn't *poach* a *chicken* any other way.

Since some predicates select specifically for complex types, some dot objects may function as disambiguators for such predicates. Consider the verb *dictate*, which has two main senses: (1) "verbalize to be recorded", and (2) "control" (possibly split into "control" with animate subjects and "serve as motivation for" with inanimate subjects). The following nouns all occur[2] as direct objects with the first sense of *dictate*:

(7.4) a. passage, story, letter, memoirs, novel
b. message, words, work, point

However, the nouns in (7.4a) are the good disambiguators (i.e. they can not be dictated in the "control" sense). The nouns in (7.4b) are ambiguous. The good disambiguators are actually dot objects of type INFO • PHYSOBJ, with *dictate* functioning as a gating predicate, which requires for the information to be given physical form.

The use of complex types in text suggests that there is an inherent asymmetry in the way dot objects are used. This asymmetry is consistent with the systematic relation between the senses, where each sense corresponds to one of the component types. For example, for the ANIMAL • FOOD nominals, the subject position tends

[1]This listing was first provided first in Pustejovsky (2005) and expanded in Rumshisky et al. (2007).

[2]The data below is taken from the British National Corpus (BNC)

to disprefer the FOOD sense, whereas in the object position, such nominals occur both with the FOOD- and the ANIMAL-selecting predicates, as well as with the gating predicates. In the object position, the FOOD selectors and the gating predicates tend to dominate:

(7.5) *chicken.n*
subject
a. ANIMAL: peck, look, wander, come, cross, follow, die
object
a. ANIMAL: count, chase, kill, shoot, slaughter, skin, pluck, sacrifice, throw
b. FOOD: eat, serve, prefer, turn, dip, stuff, carve, baste, roast, simmer
c. ANIMAL • FOOD: poach, cook

A similar asymmetry can be seen with respect to different argument positions for such dot types as PROCESS • RESULT, EVENT • PROPOSITION, etc. For example, adjectival modifiers for *construction* (PROCESS • RESULT) tend to select for RESULT, whereas the predicates that take *construction* as direct object tend to select for PROCESS. Similarly, for *allegation* (EVENT • PROPOSITION), the PROPOSITION interpretation is preferred in the object position.

(7.6) *construction.n*
object
EVENT: finance, oversee, complete, supervise, halt, permit, recommend enable, delay, stimulate
PHYSOBJ: examine, build, inaugurate, photograph
adjectival modifier
PHYSOBJ: logical, syntactic, passive, solid, all-metal, geometric, hybrid, rugged, sturdy, artificial, cultural, imaginative

(7.7) *allegation.n*
object
EVENT: face, fuel, avoid, deflect
PROPOSITION: deny, refute, counter, contain, substantiate, rebut, confirm, believe, corroborate, hear, dispute, broadcast, prove

Generic asymmetry of use (i.e. the asymmetry across all argument positions) is also a common property of some dot nominals. For example, such PROCESS • RESULT nominals as *building*, *invention*, *acquisition* show a distinct preference for one of the types in all argument positions. For *building* and *invention*, the RESULT/PHYSOBJ

interpretation is much more frequent, whereas for *acquisition*, the PROCESS/EVENT interpretation dominates the use in all argument positions. In (7.8)–(7.10) below, we list the lexical items that tend to select each component type (or the dot type itself) for these nouns in selected argument positions[3].

(7.8) *invention.n*
object
a. RESULT: produce, explain, protect, adopt, develop, combine, patent, license, display, neglect, export, exploit
b. PROCESS: welcome, avoid, stimulate, spark, trace, facilitate, demand
subject
a. RESULT: simplify, impress, consist, popularize, appear, comprise
adjectival modifier
a. RESULT: finest, original, comic, successful, British, latest, patented, brilliant

(7.9) *building.n*
object
a. PHYSOBJ: erect, demolish, construct, occupy, restore, enter, convert, design, destroy, lease, own, renovate, surround, damage, complete
b. EVENT: allow, finish, oppose, accelerate, initiate, halt, commence, stop, undertake
c. EVENT • RESULT: plan
d. EVENT, RESULT: arrange, abandon
subject
a. PHYSOBJ: house, stand, collapse, contain, survive, belong, remain, overlook, surround, fall, replace, dominate
b. EVENT: begin, continue, commence
c. EVENT • PHYSOBJ: date
d. EVENT, PHYSOBJ: accompany

(7.10) *acquisition.n*
object
a. EVENT: finance, fund, complete, announce, authorize, commence, facilitate, oversee, control, approve, undertake
b. RESULT: identify, secure, seize, store, stalk
subject
a. EVENT: occur, boost, result, strengthen, increase, depend, form, take, con-

[3]Note that for *building*, for example, *plan* selects for the complex type EVENT • RESULT in the object position, while *abandon* may select for either of the component types.

tinue, affect, result
b. RESULT: turn out, offer, comprise, bore, allow
c. EVENT • RESULT: put, increase, mean, represent, complement

Subphrasal syntactic cues (e.g. plural/singular, definite/indefinite article) are often strong indicators of the likely type selection:

(7.11) a. He stored all his new *acquisitions* here. (**plural**, RESULT)
b. The city authorized the *acquisition* of land to build the tunnel. (**singular**, EVENT)

(7.12) a. It was the most important development in radio since *the invention* of the transistor. (**definite**, EVENT)
b. *An invention* may be very beneficial, but it might also seriously undermine an existing business. (**indefinite**, RESULT)

However, the asymmetry inherent in a particular dot object may easily overrule even the strong contextual indicators. For example, *acquisition* still tends to favor the EVENT interpretation even in plural, whereas even the use with an aspectual predicate does not override the preference of *building* for the RESULT interpretation:

(7.13) a. *Acquisitions* have formed an important part of our strategy.
b. The *building* was never *completed*.

Complex types comprised by more than two component types, such as *lecture* (EVENT • (INFO • PHYSOBJ)) or *newspaper* (ORGANIZATION • (INFO • PHYSOBJ)) in (7.14) and (7.15), are also quite common. The context in which they occur may operate on any combination of the component types, including coercive operations, as in "accuse the newspaper of treason" (ORGANIZATION → HUMAN).

(7.14) *lecture.n*
object
a. EVENT: attend, organize, schedule, miss, finish, arrange, sponsor, continue, end
b. PHYSOBJ • INFO: write, follow, summarize, publish, record, illustrate, entitle, publish, understand, prepare, deliver, present
a_modifier
a. EVENT: inaugural, annual, impromptu, public, plenary, earlier, fifty-minute, weekly, regular, short, free, open
b. INFO: illustrated, introductory, stern, fascinating, anatomical, admirable,

entertaining, scientific, brilliant, feminist, good, excellent
c. EVENT, INFO: popular
d. EVENT • INFO: formal

(7.15) *newspaper.n*
object
a. PHYSOBJ: fold, open, toss, drop, stick, rip, throw, hold
b. PHYSOBJ • INFO: edit, publish, read, print, produce, ban, distribute, deliver
c. ORGANIZATION: own, launch, establish, control, tell, accuse, contact
subject
a. ORGANIZATION: report, interview, criticize, attack
a_modifier
a. (PHYS • INFO) • ORGANIZATION: daily, weekly, quarterly local, foreign, national, provincial, popular, independent, leading, oldest
b. PHYS: rolled-up, folded, yellowed, discarded
c. INFO: tabloid, filthy, conservative, right-wing, serious, English-language

7.1.2 Clustering Task

This complexity of selectional behavior makes it difficult to apply to dot objects the notion of word sense as it is used in various automatic text processing tasks. For example, multiple selection, as illustrated in (7.2), makes it impossible to resolve the classification problem of word sense disambiguation. However, as we have seen in the examples above, in many cases, it is possible to tell which type (or types) a particular individual selector prefers.

In the rest of this section, we will use the word *lunch* to illustrate how to obtain a clustering of selectors according to the type they select from the complex type. Manual inspection of the combinatorial behavior of *lunch* yields the following groupings:

(7.16) *lunch.n*
object
a. FOOD: eat, cook, enjoy, prepare, take, bring, etc.
b. EVENT: skip, attend, miss, host, cancel, etc.
adjectival_modifier
a. FOOD: light, delicious, three-course, excellent, liquid, home-cooked, half-eaten, heavy, substantial, etc.
b. EVENT: leisurely, early, annual, celebratory, official, private, weekly, etc.

The data is similar to what we have seen with respect to the noun selectors of

polysemous verbs. Here, verb selectors such as *cancel* and *attend* each have very different sets of senses, and their frequencies of occurrence do not have a similar distribution across contexts. However, with respect to the context (*lunch*, obj^{-1}), they are quite similar: they both select for the EVENT interpretation.

Clustering selector contexts according to the type they select (e.g. predicates that select for the EVENT interpretation of *lunch* vs. those that select for the FOOD interpretation) is induced by clustering *selectional equivalents* of the target noun. For nouns, we will use the term *contextual synonyms* to refer to selectional equivalents of the target.

7.1.2.1 Algorithm Description

In order to determine which type each selector activates, we proceed as follows:

1. Follow the steps described in Chapter 4, Section 4.2.1 to produce a cluster tree for the specified selector type of the target dot object.

   For example, for the target context $(t, R) = (\textit{lunch}, \mathsf{obj}^{-1})$, we would cluster all nouns that occur with the verbs that take *lunch* as direct object.

2. Select a certain number of *seed* elements from the contextual synonyms with highest contextualized similarity to the target. Trace their merges in the dendrogram, obtaining a trace sequence of clusters $C_0^i \subset ... \subset C_{n_i}^i$ for each seed i, where $C_0^i = \{i\}$, and $C_{n_i}^i$ is always the top cluster.

   (a) Sort the clusters in each trace sequence on the percent decrease in APS obtained at the next merge, and select the top-scorers.

   (b) If a cluster is among the top-scorers for several seeds, select that cluster to represent one of the senses of the target.

3. Compute the percent decrease in APS (APS derivative) for every cluster merge point. Cut the dendrogram trace at the point that has a high percent decrease in APS, so as to select the cluster obtained prior to the APS-decreasing merge.

4. For each of the target's selectors s in grammatical relation R, compute the following score for each of the chosen clusters C:

$$assoc(s, C) = \sum_{w \in C} assoc_R(s, w) \tag{7.17}$$

   The resulting score indicates how likely selector s is to pick the sense of the target associated with C. The difference between the scores obtained for different senses with a given selector indicates how strongly that selector tends to

prefer one of the senses. If the difference is small, the selector must either (1) select for the complex type itself, or (2) equally likely select for either of the component types.

Seed selection in Step 2 above may be performed automatically or an external source may be used for seed selection. We used distributional information to select seeds from the inventory of synonyms in Oxford Thesaurus of English.

7.1.2.2 Resulting Selector Assignment

We illustrate below the outcome of different stages of the algorithm using the target context $(t, R) = (\textit{lunch}, \mathsf{obj}^{-1})$, i.e. classifying verbs that accept *lunch* as direct object according to whether they activate *Event* or *Food* interpretation for *lunch*.

In this section, we use the following configuration:

(1) Grammatical relations were extracted with the Sketch Engine library.

(2) We used CP-PROD configuration setting, cf. Table 4.2.

(3) Similarity was computed over the lists of top-k selectors, with $k = 20$.

(4) All nouns that co-occurred with 5 or more of the target's selectors were chosen for clustering.

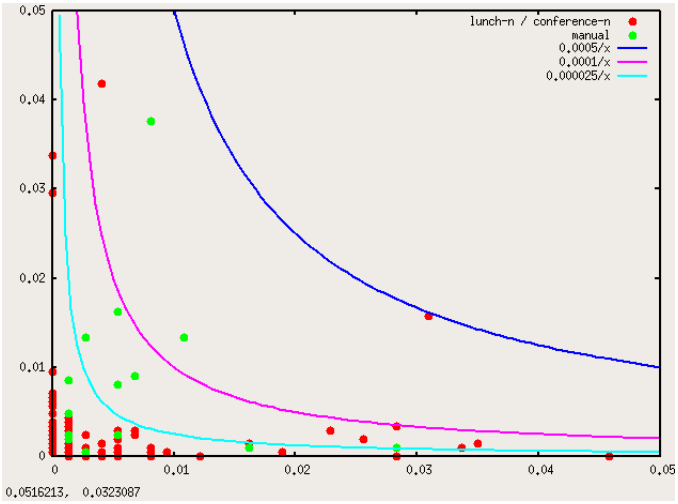

Figure 7.1: Choosing selectors for the noun pair *lunch-n/conference-n*

7.1.2.2.1 Selector lists The choice of selector lists used to compute similarity between contextual synonyms is illustrated in Tables 7.2 and 7.3. Table 7.2 shows the

top-10 selectors for the contextual synonyms of two different senses of *lunch*, as used in direct object position. The two words shown, *conference-n* and *sandwich-n*, are contextual synonyms to the EVENT and FOOD senses of *lunch-n*, respectively. Selector sets chosen for them reflect this distinction. That is, you *attend*, *hold*, *organize*, etc., *lunch* in the same way you would a *conference*, while you *eat*, *get*, *serve*, etc., *sandwiches* the same way you do *lunch*. Figure 7.1 shows the choice of selectors for the obj^{-1} context of *lunch-n/conference-n*. Correctly chosen selectors are depicted in green, with conditional probabilities for the target $P(s|Rt)$ along the x-axis, and conditional probabilities for the contextual synonym $P(s|Rw)$ along the y-axis. Note that if the incorrect selectors (in red) fall into the same equivalence class as the correct ones (in green), but have a lower association score with the contextual synonym, they will have a smaller cumulative impact during the similarity computation.

In selector lists obtained for *conference* with respect to *lunch*, there are at least 3 verbs that seem either inappropriate or incorrect: *tell* seems inapplicable with respect to taking *lunch* as direct object, and *take* and *get* seem much more likely to select for the PHYSOBJ aspect of *lunch*. Several things should be noted here. First, the values for the correctly chosen selectors cumulatively seem to insure that similarity between the selectional equivalents of the same sense is much higher than their similarity with selectional equivalent of the other sense (cf. Table 7.3). Secondly, light verbs such as *take* and *get* which are generally hard to classify correctly contribute little to the overall similarity value. And finally, the verb *tell* which initially seems to be a bizarre choice, in fact occurs in the BNC strictly with the EVENT sense of *lunch*.

7.1.2.2.2 Cluster choice A partial trace of the dendrogram obtained for (*lunch*, obj^{-1}), using the seed *conference* is shown in Table 7.4. It is easy to see that semantically very distinct words begin to cluster very early in the trace, yet most of the elements in the initial merges are clearly good contextual synonyms for the EVENT sense of *lunch*.

Figure 7.2 shows the decrease in intra-cluster APS value in the resulting sequence of merges, as well as the rate of this decrease at each merge. The final clusters that represent target's senses are selected from the clusters with highest decrease rate in intra-cluster APS. Note that the decrease rate in *inter-cluster* APS (i.e. the similarity between merged clusters) also contributes to the resulting cluster quality. In order to improve the final cluster choice, the inter-cluster and intra-cluster APS decrease values may be combined to obtain a composite indicator of the resulting cluster quality.

Consider the best clusters obtained for the contextual synonyms of the EVENT and FOOD senses of *lunch* in direct object position:

```
Cluster 6290=5702+6230:
[juice-n,cocktail-n,alcohol-n,wine-n,ale-n,brandy-n,vodka-n,champagne-n,beer-n,pint-n,
```

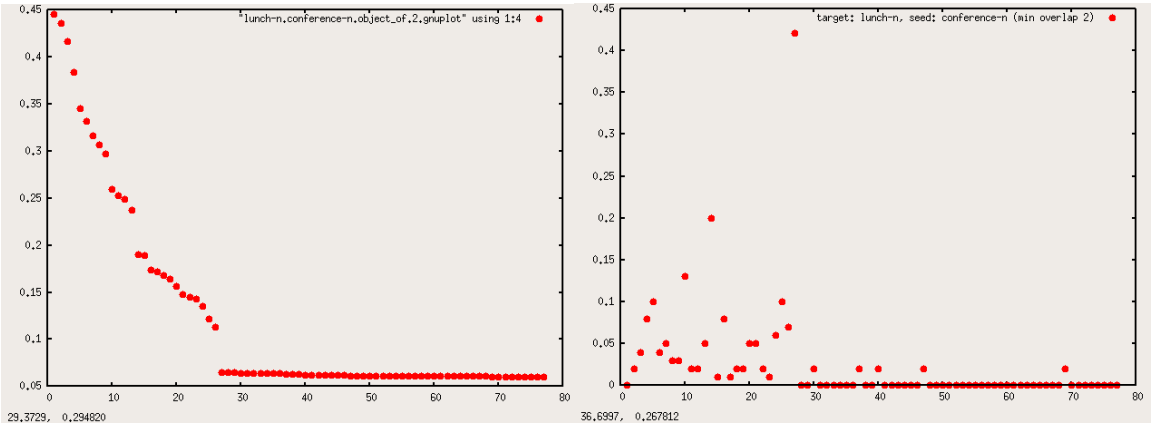

Figure 7.2: Intra-cluster APS decrease (left) and decrease rate (right) for the target *lunch*, with seed *conference*.

```
whisky-n,gin-n,sherry-n,straw-n,corn-n,liver-n,cereal-n,goose-n,vegetable-n,rice-n,
pasta-n,stuffing-n,dish-n,tomato-n,pea-n,bean-n,ham-n,turkey-n,mushroom-n,potato-n,
chicken-n,carrot-n,bacon-n,cabbage-n,nut-n,apple-n,orange-n,lettuce-n,dessert-n,chip-n,
food-n,snack-n,buffet-n,steak-n,salad-n,sandwich-n,dinner-n,meal-n,lunch-n,breakfast-n,
supper-n,beef-n,sweet-n,crisp-n,chop-n,sausage-n,pizza-n,meat-n,chocolate-n,banana-n,
spaghetti-n,yoghurt-n,ice-cream-n,doughnut-n,mint-n,honey-n,jam-n,soup-n,toast-n,tea-n,
coffee-n,bread-n,cheese-n,cake-n,curry-n,bun-n,biscuit-n,pudding-n,marmalade-n,jelly-n,
pie-n,porridge-n,tart-n,pastry-n,stew-n,sauce-n,hay-n,butter-n,roll-n,cream-n]

Cluster 6347=5673+6299:
[tournament-n,contest-n,outing-n,barbecue-n,exhibition-n,festival-n,hearing-n,summit-n,
talk-n,ballot-n,election-n,referendum-n,disco-n,congress-n,inquest-n,fair-n,ceremony-n,
reunion-n,rally-n,meeting-n,conference-n,seminar-n,parade-n,rehearsal-n,dance-n,funeral-n,
clinic-n,feast-n,celebration-n,session-n,workshop-n,demonstration-n,concert-n,briefing-n,
lecture-n,reception-n,banquet-n,luncheon-n,wedding-n,gathering-n,event-n,procession-n]
```

For comparison, we ran Pantel's CBC algorithm (Pantel, 2003) on the same corpus, that is, the 100M word British National Corpus. The first m highest-ranking elements of the clusters obtained for the word *lunch* are shown below, where m is the size of the cluster obtained for the corresponding sense by our algorithm.

```
{N357 beer, wine, drink}

beer, wine, drink, food, cigarette, Tobacco, beverage, grocery, booze, fag, "sparkling wine",
"Scotch whisky", caffeine, cigar, Soda, liquor, toothpaste, cornflake, snuff, confectionery,
brew, dram, titbit, incense, alcohol, pint, sweet, pizza, stout, wee, "fish and chips", Pasta,
hamburger,condom, cracker, bottle, squash, fizz, pee, morsel, Chardonnay, sausage, painkiller,
pints, dynamite, concoction, aspirin, vinegar, potion, lunch, cocaine, yoghurt, pie, curry,
gallon, dinner, warmer, spice, feed, beef, balm, bitter, yogurt, nappy, toothbrush, cannabis,
LOBSTER, contraceptive, steak, margarine, tonic, Oyster, tumbler, Mead, tankard, glass, perfume,
flask, "Christmas card", Rice, medicine, garlic, dish, butter, grouse, apple, feller, bookcase,
fish

{N270 trip, visit, tour}
```

```
trip, visit, tour, holiday, journey, expedition, excursion, cruise, jaunt, voyage,
trek, honeymoon, outing, pilgrimage, reunions, "social event", detour, flight, sortie,
Sightseeing, travel, sojourn, "long haul", fling, mission, Odyssey, escapade, backpacking,
"sick leave", "The drive", "growing season", walkabout, convalescence, foray, safari,
stopover, regatta, ascent, roadshow errand, countdown, gestation
```

For the target context $(t, R) = (lunch, \mathsf{obj}^{-1})$, our algorithm seems to give comparable or better results. The resulting clusters are more homogeneous and contain fewer spurious elements.

7.1.2.2.3 Selector assignment Table 7.5 shows the soft selector assignment obtained for $(lunch, \mathsf{obj}^{-1})$ using the above clusters as described in step 4 of Section 7.1.2.1. Notice that the selector sets for both senses are quite heterogeneous, but the assigned selector/sense pairings seem to be accurate in the majority of cases. The incorrect assignment often produces a low confidence rating, as with *skip*, for example.

The accuracy of assignment can sometimes be difficult to judge without looking at the actual usage. For example, *hold* gets assigned the highest association score with the EVENT sense. This may appear inaccurate, since *hold* is quite polysemous and one of its senses selects for PHYSOBJ. However, in all occurrences of *lunch* in the BNC, *hold* is indeed found with the EVENT interpretation, actually confirming the accuracy of the assigned scoring.

7.2 Modifier-Based Disambiguation of NPs

In this section, we come back to verbal polysemy, and consider additional factors that determine which sense of the verb is activated by the relevant argument.

Our method for verb sense induction presented in Ch. 4 assumes that in most cases semantic load of the argument NP will be carried by the head noun. Obviously, in many cases semantic load is carried by other elements. This applies especially to semantically weak head nouns.

For example, consider the word *position* whose meaning is so underspecified that it almost always requires a modifier in order to be disambiguated. Thus, in (7.18), the adjectival modifier is effectively the sole factor determining both the meaning of the word *position* and the interpretation assigned to *assume*.

(7.18) a. He instantly assumed a kneeling position.
b. He instantly assumed a managerial position.
c. He instantly assumed an antiracist position.

We used the RASP parser to extract from the BNC the words that occur in the ncmod

relation with *position*,[4] and ranked them according to their association scores with *position*, using the log-factor adjusted MI. The top-scoring modifiers of *position* under this scoring scheme were *sitting*, *predicative*, and *dominant*. Using the dendrogram obtained for the ncmod relation for *position*, we can sort the clusters whose selector lists include a given modifier, so that the cluster in which that modifier has the highest average MI is placed at the top. For the modifiers above, this method places at the top the clusters with the following selector lists[5]:

sitting: stooped-j 11.4, kneeling-j 11.3, recumbent-j 11.0, seated-j 10.1, commanding-j 8.5, standing-j 7.5, ...
cluster: [*figure-n posture-n*]

predicative: attributive-j 14.1, predicative-j 13.7, postnominal-j 13.2, clausal-j 13.1, predicate-n 10.0, postverbal-j 9.1, syntactic-j 8.8, prenominal-j 8.3, ordinal-j 6.8, adjectival-j 6.1, ...
cluster: [*construction-n adjective-n*]

dominant: interactionist-n 11.4 marxist-j 10.3, pluralist-n 10.2, philosophical-j 8.8, popperian-j 8.7, antiracist-j 7.8, phenomenological-j 7.5, kantian-j 7.4, structuralist-j 7.4, essentialist-j 7.3, functionalist-j 7.2, dominant-j 7.1, holist-j 6.9, doctrinal-j 6.6, materialist-n 6.4, theoretical-j 5.4, ideological-j 5.1, ...
cluster: [*conception-n perspective-n critique-n*]

Notice that the phrase "dominant position" is actually ambiguous between the *point of view* sense of *position* and the *relative standing* sense. The second cluster in the sorted list for *dominant* identifies the other sense:

dominant: monopolistic-j 8.1, leading-j 8.0, competing-j 7.6, respected-j 6.1, rival-j 6.0, monopoly-n 5.5, established-j 5.3, dominant-j 5.2, competitive-j 4.6, well-established-j 4.5, ...
cluster: [*manufacturer-n firm-n producer-n provider-n supplier-n*]

The resulting heterogeneous selector sets could be used to improve the resolution of lexical ambiguity in statistical machine translation. For example, in the two sentences given in (7.19), the word *position* has the same interpretation.

(7.19) a. He instantly assumed a kneeling position.
b. He instantly assumed a stooped position.

However, state-of-the-art statistical machine translation engines do not seem to always recognize this fact. For example, for Russian, Google Translate gives an appropriate translation in case of "kneeling", but the word "stooped" is not translated at all, and the word "assume" is not translated appropriately in either case:[6]

[4] In RASP, ncmod is the relation between the noun and its non-clausal modifier.

[5] MI values for each selector, averaged across all elements of the cluster, are given next to the POS-marked lemma

[6] Google Translate is available at `http://translate.google.com`. Cyrillics are transliterated; translations for the word *position* are underlined in both sentences.

(7.20) a. He instantly assumed a kneeling position.
On mgnovenno priobrel polozhenie dlya strel'by s kolena. (Google)

b. He instantly assumed a stooped position.
On mgnovenno priobrel stooped pozitsii. (Google)

The likely reason for this is that the word *stooped* is not frequent enough for the parallel corpora to provide reliable n-gram statistics. However, in the clusters our method produces for the word *position*, the words *stooped* and *kneeling* have high association scores with same ncmod-induced cluster. Using the corresponding cluster to obtain n-gram statistics for *stooped* would improve the chances of obtaining an appropriate translation for both *position* and *assume* in such cases.

7.3 Summary

In this chapter, we have examined some possible applications of the proposed clustering method. Despite the peculiar selectional behavior of dot objects, such as the multiple selection phenomena or the presence of selectors specific to the dot-type itself, it seems possible to derive automatically sets of selectors for each component type using our clustering method. It seems also that this clustering method may be applied successfully to other cases, such as the disambiguation of full NPs with semantically weak head nouns. There is clearly a number of sense detection tasks to which this method can be applied, and these tasks should be investigated in the future.

Dot type	Example
Action • Proposition	promise, allegation, lie, charge
State • Proposition	belief
Attribute • Value	temperature, weight, height, tension, strength
Event • (Info • PhysObj)	lecture, play, seminar, exam, quiz, test
Event • (Info • Sound)	concert, sonata, symphony, song
Event • PhysObj	lunch, breakfast, dinner, tea
Info • PhysObj	article, book, CD, DVD, dictionary, diary, email, essay, letter, novel, paper
Organization • (Info • PhysObj)	newspaper, magazine, journal
Organization • Loc • HumanGroup	university, city
Event • Location • HumanGroup	class
Aperture • PhysObj	door, window
Process • Result	construction, imitation, portrayal, reference, decoration, display documentation, drawing, enclosure, entry, instruction, invention, simulation, illustration, agreement, approval, recognition, damage, compensation, contribution, discount, donation, acquisition, deduction, endowment, classification, purchase
Producer • Product	Honda, IBM, BMW
Tree • Fruit / Tree • Wood	apple, orange, coffee / oak, elm, pine
Animal • Food	anchovy, catfish, chicken, eel, herring, lamb, octopus, rabbit, squid, trout
Container • Contents	bottle, bucket, carton, crate, cup, flask, keg, pot, spoon

Table 7.1: Some examples of dot objects of different complex types, as well as "pseudo-dots" that exhibit dot-like behavior due to coercion.

	lunch-n		*sandwich-n*			*lunch-n*		*conference-n*	
	count	$P(v\|Rn)$	count	$P(v\|Rn)$		count	$P(v\|Rn)$	count	$P(v\|Rn)$
eat	93	.1253	93	.2035	attend	15	.0202	263	.1251
take	48	.0647	30	.0656	hold	10	.0135	379	.1803
get	40	.0539	25	.0547	give	23	.0310	33	.0157
make	17	.0229	56	.1225	tell	2	.0027	285	.1356
want	19	.0256	17	.0372	organize	6	.0081	79	.0376
bring	21	.0283	13	.0284	take	48	.0647	6	.0029
finish	21	.0283	8	.0175	call	3	.0040	88	.0419
buy	14	.0189	12	.0263	arrange	8	.0108	28	.0133
prepare	21	.0283	7	.0153	get	40	.0539	4	.0019
serve	42	.0566	3	.0066	bring	21	.0283	7	.0033

Table 7.2: Top 10 selectors for the noun pairs *lunch-n/sandwich-n* and *lunch-n/conference-n* in direct object position.

	lunch-n	conference-n	fair-n
attend-v	0.020	0.125	0.066
hold-v	0.013	0.180	0.264
tell-v	0.003	0.136	0.022
organise-v	0.008	0.038	0.011
arrange-v	0.011	0.002	0.011
host-v	0.005	0.016	0.033
follow-v	0.007	0.009	0.011
organize-v	0.003	0.013	0.011
csim(conference, fair)			0.319

	lunch-n	sandwich-n	fair-n
get-v	0.054	0.055	0.022
csim(sandwich, fair)			0.022

Table 7.3: Similarity computation for contextual synonyms of two senses of *lunch*. Association scores for the intersection of top-k selector lists are shown for: A. (left) *conference* and *fair*. B. (right) *sandwich* and *fair*.

Step	*Inter-cluster APS*	*Intra-cluster APS*	*APS % decrease*	*Resulting cluster*
1	0.445	0.445	0.00	[conference-n] [seminar-n]
2	0.430	0.435	0.02	[meeting-n] [conference-n seminar-n]
3	0.397	0.416	0.04	[rally-n] [meeting-n conference-n seminar-n]
4	0.342	0.387	0.07	[reunion-n] [rally-n meeting-n conference-n seminar-n]
5	0.314	0.363	0.06	[ceremony-n] [reunion-n rally-n meeting-n conference-n seminar-n]
6	0.295	0.332	0.09	[inquest-n fair-n] [ceremony-n reunion-n rally-n meeting-n conference-n seminar-n]
7	0.267	0.318	0.04	[congress-n] [inquest-n fair-n ceremony-n reunion-n rally-n meeting-n conference-n seminar-n]
8	0.264	0.307	0.03	[disco-n] [congress-n inquest-n fair-n ceremony-n reunion-n rally-n meeting-n conference-n seminar-n]
9	0.246	0.280	0.09	[talk-n ballot-n election-n referendum-n] [disco-n congress-n inquest-n fair-n ceremony-n reunion-n rally-n meeting-n conference-n seminar-n]
10	0.223	0.272	0.03	[summit-n] [talk-n ballot-n election-n referendum-n disco-n congress-n inquest-n fair-n ceremony-n reunion-n rally-n meeting-n conference-n seminar-n]
11	0.216	0.265	0.03	[hearing-n] [summit-n talk-n ballot-n election-n referendum-n disco-n congress-n inquest-n fair-n ceremony-n reunion-n rally-n meeting-n conference-n seminar-n]
12	0.197	0.224	0.15	[hearing-n summit-n talk-n ballot-n election-n referendum-n disco-n congress-n inquest-n fair-n ceremony-n reunion-n rally-n meeting-n conference-n seminar-n] [parade-n rehearsal-n wedding-n funeral-n clinic-n feast-n celebration-n session-n workshop-n demonstration-n concert-n briefing-n lecture-n reception-n banquet-n luncheon-n]
...				...

Table 7.4: Dendrogram trace for the target *lunch*, seed *conference*.

Selector	FOOD	EVENT	Assigned Type	Confidence	Selector	FOOD	EVENT	Assigned Type	Confidence
eat-v	0.089	0.002	food	.087	cancel-v	0.000	0.003	event	.003
cook-v	0.024	0.003	food	.021	organise-v	0.000	0.034	event	.034
serve-v	0.024	0.002	food	.022	include-v	0.013	0.011	food	.002
skip-v	0.002	0.000	food	.002	order-v	0.008	0.001	food	.007
finish-v	0.009	0.002	food	.007	grab-v	0.000	0.000	food	.000
enjoy-v	0.006	0.016	event	.010	give-v	0.010	0.045	event	.035
prepare-v	0.009	0.004	food	.006	spoil-v	0.000	0.000	food	.000
attend-v	0.001	0.100	event	.098	share-v	0.004	0.002	food	.002
miss-v	0.001	0.002	event	.001	hold-v	0.004	0.157	event	.153
take-v	0.023	0.007	food	.016	pack-v	0.000	0.000	food	.000
provide-v	0.007	0.010	event	.003	appreciate-v	0.000	0.000	food	.000
get-v	0.064	0.014	food	.050	like-v	0.032	0.004	food	.028
bring-v	0.011	0.003	food	.008	offer-v	0.006	0.003	food	.003
buy-v	0.023	0.000	food	.023	plan-v	0.000	0.013	event	.013
arrange-v	0.002	0.019	event	.017	supply-v	0.001	0.000	food	.001
want-v	0.035	0.003	food	.032	make-v	0.083	0.016	food	.067
host-v	0.000	0.010	event	.010	organize-v	0.000	0.011	event	.011

Table 7.5: Selector assignment scores for (*lunch*, obj^{-1}). A-score$(s) = \sum_{e \in C} P(s|e)$, where C is a cluster of selectional equivalents (contextual synonyms) of the target word corresponding to one of its senses. Confidence is computed as the raw difference between the two A-score values.

Chapter 8

Conclusions

In this thesis, we have examined the problem of verbal polysemy, with a particular focus on sense distinctions that are detected based solely on the semantics of the verb's arguments. We have explored the issues involved in identifying such sense distinctions, both manually and automatically. We provided an analysis of how human speakers deal with this problem, and an automatic algorithm aimed at modeling such phenomena. We have also looked in some detail at the nature of sense definition and the perils of sense inventory construction for polysemous verbs.

Evaluation framework We have argued that the standard framework currently used in the field to evaluate the success of word sense detection systems suffers from a lack of discriminatory power. Computing the overall performance accuracy of such systems does not provide accurate or useful analysis of their successes and limitations. We therefore proposed a new way of looking at this problem. We argue that an effective evaluation must allow one to examine the types of sense distinctions successfully detected by the system and in particular, to evaluate how well the system recognizes various factors that contribute to sense differentiation.

The ideal solution would be to create a data set that contains a sufficient number of instances for each type of sense distinction we wish to be able to detect, specifying the factors relevant for the disambiguation in each case, as well as the type of sense distinction involved. Modifying the CPA-style annotation scheme to include such information, perhaps, would yield such a data set. However, we wanted to explore the general possibility of separating out sense distinctions linked to a particular set of contextual factors.

Contributions We created a semantically annotated data set targeting one of the least studied factors that contribute to sense disambiguation for the verbs, namely, the semantics of the arguments. More specifically, we focused on sense distinctions

that can be detected by looking at the semantics of a single argument. The outcome of this effort is two-fold. Firstly, it allowed us to examine how the speakers deal with such verbal ambiguities. Secondly, it provided the testing data for any algorithm that would handle these ambiguities.

We have also presented a clustering algorithm that allows us to produce sets of words selectionally similar to a given sense of the verb and induce clusters of arguments activating that sense in a fully unsupervised setting. We avoid the common computational pitfalls in distributional similarity-based clustering by computing clusters of short contextualized vectors. Contextualizing the representation of the verb's selectional equivalents to the target context insures that we capture the verb's selectional properties specific to that context.

Applications of the technique The output produced by the clustering algorithm can be used in a number of ways, in tasks related to sense disambiguation. The derived information about selectional properties of different senses of the target word can serve to improve the overall performance of a complete WSD or WSI system. We have discussed how it can be applied to the resolution of nominal polysemy in Chapter 7. Other obvious candidates include various parsing tasks such as PP-attachment or NP-parsing (when the same technique is applied to polysemous nouns).

It can also provide powerful enhancements to the lexicographic analysis tools that facilitate sense definition. For example, it can be used to create contextualized clusters of collocates in an application such as the Sketch Engine (Kilgarriff et al., 2004). In fact, examining the induced sets of selectional equivalents often reveals unexpected relationships between verbs that accept similar arguments in a given argument position. The discovered selectional equivalence relations are often impossible to predict by inspecting the data with traditional methods. This suggests that the presented technique for automated analysis of selectional properties can also be viewed as a tool for a more focused empirical study of the data. In particular, it may serve to enrich the initial models of the data – the theoretical models that are often limited to using the introspective intuition and targeted corpus studies.

Future work The algorithm we presented can be extended or improved in a number of ways. In a standard WSD task, the verb's sense gets activated by a combination of selectors in different argument positions, so it is clearly desirable to use an extended set of grammatical relations instead of single relation inverses. However, an attempt to use the context extending beyond a simple binary relation with this method quickly runs into a sparsity problem. In other words, without an appropriate generalization mechanism, it is infeasible to create, for example, contextualized clusters of (subject, object) pairs for the verb – rather than merely clusters of semantically diverse direct objects.

As a possible solution, one can envision creating a many-to-one mapping between the clusters of selectional equivalents produced for two different argument positions. The mapping can be constructed, for example, based on the available co-occurrence statistics for the words in these positions, with each actual subject/object pair observed in a corpus increasing the likelihood of a link between the corresponding clusters. The resulting mapping would effectively make it possible to combine the association scores from selectors in different argument positions.

This clustering procedure we defined also allows one to seed the clusters manually, as it is done in some thesaurus construction algorithms (e.g. Roark and Charniak, 1998). The dendrogram produced by the algorithm can be partitioned as suggested in Ch. 7, by manually specifying several selectional equivalents (or, equivalently, the corresponding selectors) for each sense.

Lessons from the targeted sense annotation Our motivation for creating the sense-annotated data set was also to investigate the feasibility of evaluating separately the contribution of a particular type of context feature to sense disambiguation. Our annotation effort has demonstrated that such separation is possible. The resulting data set can be used to evaluate how well any given WSD or WSI system handles sense distinctions that are dependent on the semantics of the arguments.

The goal of annotating context features contributing to disambiguation can also be accomplished by incorporating this information into sense inventory specification. This can be accomplished by extending the existing annotation schemes to include the information about sense relations and context elements that activate each senses. For example, CPA already requires the lexicographer to specify the full range of context features for each pattern of use. Only a slight modification of the CPA annotation scheme would allow one to identify the sense-distinguishing elements for each pattern. Identifying relations between different senses of the same word can also be done when a sense inventory for that word is compiled. Such annotation does not need to be unnecessarily complex, but it will require developing some robust generalizations about sense relations.

We have argued that the only instances kept in the data set should be the ones for which the disambiguation can be performed reliably by human speakers. This seemingly controversial suggestion is motivated by the fact that the examples which are unclear or difficult to disambiguate introduce noise into the data set without contributing any substantial information about the corresponding sense distinctions.

In our annotation task, we have explicitly instructed the annotators to throw out the examples that are unclear for any reason. However, our annotation scheme did not allow the annotators to specify the reasons for discarding each case. In the course of this work, it became evident that it would be easy and interesting to collect this information during the annotation. In the annotation effort that continues this work,

the annotator is given the following options:

(i) no sense seems to fit (*sense not in the inventory*)

(ii) more than one sense seems to fit (*boundary case*)

(iii) impossible to establish from context which sense was used (*insufficient context*)

(iv) creative or metaphoric use of a sense

A new annotation effort currently under way applies a more fine-grained analysis to the last group of examples. The Generative Lexicon Markup Language (GLML) initiative focuses on annotating compositional processes at work in argument selection, including the mechanisms that license creative use, such as the type-shifting operations (Pustejovsky et al., 2008b; Pustejovsky et al., 2008a; Pustejovsky et al., 2009).

Appendices

Appendix A

Resources

A.1 Corpora, Parsers, and Lexical Resources

British National Corpus (BNC)

The British National Corpus (BNC) (1994) is a balanced synchronic British English text collection that contains 100 million words from a variety of sources, including written and spoken language.

Robust Accurate Statistical Parsing (RASP)

Robust Accurate Statistical Parsing (RASP) system (Briscoe and Carroll, 2002) tokenizes, POS-tags, and lemmatizes text, generating a forest of full parse trees for each sentence and associating a probability with each parse. For each parse, RASP produces a set of grammatical relations, specifying the relation type, the headword, and the dependent element. All our computations are performed over the single top-ranked tree for the sentences where a full parse was successfully obtained. Some of the grammatical relations identified by RASP are shown in 1.1.

(1.1) **subjects**: ncsubj, clausal (csubj, xsubj)
objects: dobj, iobj, clausal complement
modifiers: adverbs, modifiers of event nominals

Oxford Thesaurus of English (OTE)

Oxford Thesaurus of English (OTE) (2000) contains 16,000 entries. Synonyms, alternative and opposing words are specified for each of the senses of each entry word.

A.2 The Sketch Engine

The Sketch Engine system implements a fully automated word sketch and thesaurus construction process for any specified corpus. The thesaurus entry for each lemma is constructed by clustering lexical items occurring in a particular grammatical relation (GR) with a target lemma. Below, we describe the association and distance metrics used for word sketch and thesaurus construction for a given lemma.

The Sketch Engine includes a comprehensive concordancing system that implements full service corpus query processing and collocate statistics computation. The sections below detail the provided services provided, including collocate statistics, and word sketch and thesaurus construction.

Collocate statistics

The system provides a number of association metrics for any pair of collocates, including:

(1) T-score

(2) Mutual Information (MI) (Church and Hanks, 1990)

(3) MI^3 (Oakes, 1998)

(4) Log likelihood (Dunning, 1993)

(5) Min. sensitivity (Pedersen, 1998)

(6) Salience (Kilgarriff and Tugwell, 2001)

Word Sketches

A word sketch for a target lemma consists of a set of grammatical relations the lemma participates in, with a set of significant collocates identified for each grammatical relation. Lexical items that occur in a given grammatical relation with the target lemma are extracted by running a set of KWIC queries defined for that grammatical relation on an indexed corpus.

Grammatical relations A set of grammatical relations is pre-defined for each part of speech. Collocates that occur in that grammatical relation with the target lemma and that receive a high *association score*, based on a given corpus, are displayed for those relations.

All relations defined in the Word Sketch Engine specification are scored according to how often the target lemma participates in that relation. If the target lemma occurs in a given relation only as frequently as is the average for its POS category, the *relation association score* (RAScore) is 1. *Relation association score* below 1 implies that this relation is less typical of the target lemma, and *relation association score* implies that this relation is more typical of the target lemma than is the average for its POS category. *Relation association score* is essentially a likelihood ratio that is computed as follows, following the notation in Lin (1998):

$$\text{RAScore } (w_1, R) = \frac{P_{MLE}(w_1|R)}{P_{MLE}(w_1)} = \frac{||w_1,R,*||\cdot||*,*,*||}{||*,R,*||\cdot||w_1,*,*||}$$

In addition to the pre-defined relations, significant collocates are displayed for the top-N high-scorers amongst the other extracted relations.

Association scores The *association score* (AScore) is computed for two lemmas and a particular grammatical relation. It is used to determine the significant collocates of the target lemma for that grammatical relation. The metric currently used by the Sketch Engine is, as described in Kilgarriff and Tugwell (2001):

$$\text{AScore } (w_1, R, w_2) = \text{Pointwise MI } (w_1, R, w_2) * log(freq(w_1, R, w_2) + 1)$$

Following the notation in Lin (1998):

$$\text{AScore } (w_1, R, w_2) = \log \frac{||w_1,R,w_2||\cdot||*,*,*||}{||w_1,R,*||\cdot||*,*,w_2||} \cdot \log(||w_1, R, w_2|| + 1)$$

Thesaurus Construction

A second-order distance metric is used to identify semantically similar lexical items for the purposes of (1) constructing a thesaurus for the target lemma and (2) identifying semantic clusters amongst lexical items occuring in a given grammatical relation with the target lemma.

Distance Metric The *distance metric* (Dist) used is based on the *wordsketch difference* between two target lemmas. Here is how it is computed:

(1) Relations defined in the Sketch Engine specification for the particular corpus markup are extracted for each lemma.

(2) Word Sketch Overlap. Word sketches for the two target lemmas are considered to overlap the tuples (w_1, R_i, c_k) and (w_2, R_i, c_k) are extracted for the two lemmas w_1 and w_2. That is, word sketches overlap if:

   (i) the same relation is extracted for both target lemmas;

   (ii) both target lemmas occur with the same collocate in that relation;

   (iii) that collocate occurs more than a given number of times (currently, twice) in that relation with each target lemma;

   (iv) the *association score* of that collocate for both lemmas is greater than zero.

(3) Relation tuples extracted for w_1 and w_2 with non-matching collocate/relation pairs are considered non-overlapping. The same frequency threshold and condition on the association score is applied, i.e. tuples with frequency < 2 or association score < 0 are not considered.

(4) If a parcitular collocate c occurs in a given relation with $> 10,000$ lemmas, it is not considered in computing thesaurus distance

Let $(w_1, R_i, c_m, \mathit{freq}_1, \mathit{AScore}_1)$ and $(w_2, R_j, c_k, \mathit{freq}_2, \mathit{AScore}_2)$ be two relation tuples extracted for lemmas w_1 and w_2, with the corrponding frequencies and association scores. If $\mathit{freq}_1 > 1$, $\mathit{freq}_2 > 1$, $\mathit{AScore}_1 > 0$, $\mathit{AScore}_2 > 0$, and $R_i = R_j$ and $c_m = c_k$, we have an instance of word sketch overlap. *Distance metric* is computed by taking an adjusted sum of association scores for each overlapping tuple pair and dividing it by the sum of association scores of all relation tuples extracted for the two lemmas.

$$\mathit{Dist}(w_1, w_2) = \frac{\sum_{(\mathit{tuple}_i, \mathit{tuple}_j) \in \{\mathit{tuples}_{w_1} \cap \mathit{tuples}_{w_2}\}} \mathit{AScore}_i + \mathit{AScore}_j - (\mathit{AScore}_i - \mathit{AScore}_j)^2/50}{\sum_{\mathit{tuple}_i \in \{\mathit{tuples}_{w_1} \cup \mathit{tuples}_{w_2}\}} \mathit{AScore}_i}$$

The above formula differs from the corresponding definition in Lin (1998) in that the sum of scores of overlapping tuples is reduced by a fraction of the square of difference between the two scores. This is done in order to additionally penalize cases when the difference in association scores of the overlapping pairs is too great.

Thesaurus For the purposes of constructing thesaurus for a particular lemma, all lexical items that fall within a specified distance from that lemma (as measured by the above distance metric) are identified. The identified lexical items are further grouped according to the distance between each pair of items.

Collocate Clustering

Collocates of each lemma are also clustered within the word sketch constructed for that lemma. For each grammatical relation, the first N collocates with the highest *association score* are identified. N is three or four times the number of collocates specified in the web form (e.g., 75 for the default value of 25). Starting from the the highest scoring collocate on the list, the remaining collocates that fall within the specified distance according to the *distance metric* are grouped together in the cluster associated with the target collocate (the default distance is 0.15) . Then the procedure is repeated for the next ungrouped collocate on the list. Association scores displayed for the cluster itself are the association scores for the highest-scoring collocate in the cluster.

Appendix B

Annotation Guidelines

The general annotation instructions as presented to the annotators are shown below in Section B.1. Sense inventories and specific instructions for each verb are are given in B.2.

B.1 General Instructions

Word Sense Disambiguation for Polysemous Verbs

On the next page, you will see a list of target verbs, with a particular argument position specified for each verb. For each target verb, you will be presented with a set of sentences. Your task is to mark each sentence according to the sense in which the target verb is used in that sentence. You will be given a list of senses to choose from. The target verb will be highlighted.

Note that the list of senses for each verb is not exhaustive, rather it reflects the senses that can be distinguished based on the semantics of the argument in the specified position. The relevant argument will also be highlighted.

If you are unable to choose the appropriate sense you may mark a sentence as "N/A". Please only do so if you are certain that no valid choice can be made. Feel free to check pre-annotated data when in doubt about a particular sense.

When to mark a sentence "N/A":

1. No sense seems to fit.
2. More than one sense seems to fit.
3. There is an ambiguity: it is unclear which sense was used and impossible to tell from context.
4. There is a misparse: the target grammatical relation is detected incorrectly.

For example, the target grammatical relation may not be present in the clause at all, or be filled by a clausal argument:

> "He could not admit the house was robbed." (admit, direct object)
> "These roles are ordinarily assumed to be interchangeable." (assume, direct object)
> "The choice which fell to him was hard." (fall, subject)

Other examples of misparses include adjectives and nouns mistagged as target verbs:

> "editing job" (edit), "firing squad" (fire), "firing policies" (fire), etc.

Use your discretion when deciding whether to keep or throw out metaphoric uses of any sense.

B.2 Verb-Specific Instructions and Sense Inventories

Sense inventory for *absorb, dobj*:

1. absorb substance
   e.g. *salt, ink, water, chemicals*

2. absorb energy or impact
   e.g. *light, sound, radiation, blow, impact*

3. consume a resource, such as time or money
   e.g. *money, time, energy*

4. bear the cost of; take on an expense
   e.g. *losses, tax cuts*

5. learn or incorporate skill or information
   e.g. *values, atmosphere, information*

6. preoccupy (for a person to be preoccupied or be immersed into something)
   e.g. *people being absorbed or immersed into something that interests them*

7. take in or assimilate, making part of a whole or a group
   e.g. *refugees, immigrants being absorbed into communities; regions absorbed by countries*

Sense inventory for *acquire, dobj*:

1. take on certain characteristics
   e.g. *importance, meaning; also: reputation*

2. learn
   e.g. *language, manners, knowledge, skill*

3. purchase or become the owner of property
   e.g. *land, stocks, business*

4. become associated with something, often newly brought into being
   e.g. *cities acquiring new jobs*

Sense inventory for *admit, dobj*:

1. acknowledge the truth or reality of
   e.g. *defeat, inconsistency, offence*

2. grant entry to or allow into a community
   e.g. *patients, students; also: surface admitting water*

   Sentential arguments (e.g. that-Clauses) and other cases when there is no direct object should be considered a misparse and marked N/A.

Sense inventory for *assume, dobj*:

1. come to have a characteristic or quality
   e.g. *importance*

2. assume title, responsibility or control
   e.g. *office*

3. presuppose something to exist; take for granted
   e.g. *knowledge, validity, connection*

   Small clauses, infinitivals, etc. should be considered a misparse and marked N/A: e.g. 'Teachers are assumed biased/to be biased'.

Sense inventory for *claim, dobj*:

1. claim the truth of
   e.g. *legitimacy, high interest rate*

2. come in possession of or claim property you are entitled to
   e.g. *suitcase, inheritance; also: rights, priviledge*

3. achieve or obtain something
   e.g. *honours, victory*

   Clausal arguments should be considered a misparse and marked N/A: e.g. 'She claimed her client was not present'.

Sense inventory for *conclude, dobj*:

1. finish
   e.g. *activities, visit, lecture*

2. reach an agreement
   e.g. *treaty, deal, agreement*

   Faux passive constructions should be considered a misparse and marked N/A: e.g. 'Committee concluded by reminding..'

Sense inventory for *cut, dobj*:

1. reduce or lessen
   e.g. *supplies, costs, debt*

2. remove or stop
   e.g. *program, course*

3. make an incision or separate pieces of
   e.g. *rope, cake, hair*

4. cut out a form or a shape
   - an object generated as a result of cutting activity, e.g. length of fabric

   Phrasal verbs should be considered a misparse and marked N/A: e.g. 'They have their work cut out for them'. Adjectival uses may be preserved where appropriate, e.g. 'Cut flowers last longer'.

Sense inventory for *deny, dobj*:

1. refuse to grant something
   e.g. *access, visa*

2. state or maintain that something is untrue
   e.g. *allegations, reports, importance, allegations*

3. reject; refuse to acknowledge something
   e.g. *culture*

   Ditransitives should be considered a misparse and marked N/A: e.g. 'Their aim is to deny the batsman room to play'.

Sense inventory for *dictate, dobj*:

1. verbalize to be recorded
   e.g. *letter*

2. determine the character of or serve as motivation for
   e.g. *policy, tactics*

Sense inventory for *drive, dobj*:

1. operate a vehicle controlling its motion
   e.g. *car, truck, van*

2. travel in a vehicle a certain distance
   e.g. *a number of miles, yards, kilometers*

3. transport something or someone
   e.g. *giving a lift to a person or driving an object somewhere*

4. provide power for or physically move a mechanism
   e.g. *steam driving the engine*

5. force a vessel to move in a direction
   e.g. *winds driving a yacht on to the rocks*

6. force adversary to leave
   e.g. *competitors away, enemy off the battlefield*

7. physically urge animal to go
   e.g. *cattle, horses*

8. cause or force something or someone into a state or activity
   e.g. *drive people to madness, to despair*

9. push a sharp object into another object
   e.g. *nail, stake*

10. strike or throw an object of play
    e.g. *ball, pack*

11. motivate the progress of
    e.g. *market, research*

    Idiomatic expressions such as 'Drive a hard bargain' should be considered a misparse and marked N/A.

Sense inventory for *edit, dobj*:

1. make changes to the text; modify content; make corrections
   e.g. *text, line, file*

2. supervise publication
   e.g. *newspaper, volume, collection*

   Cases such as 'The orangutan : its biology and conservation edited by L.' should be considered a misparse and marked N/A. Adjective and nominal uses should be considered a misparse and marked N/A: e.g. 'editing jobs', 'make editing a time-consuming exercise'.

Sense inventory for *enjoy, dobj*:

1. like doing something; appreciate something
   e.g. *play, view, skiing, vacation*

2. have or possess something
   e.g. *status, success*

   Like or appreciate: e.g. taste, food, dancing, outdoors, barbecues, thrills, countryside, scenery, spectacle, sensation, school, practice, trip, holiday, visit (coerced events). Have or possess: e.g. protection, health, security, monopoly, comfort, lifestyle, notoriety, acclaim, fame, independence, autonomy, liberty, prosperity, wealth, stability (states). If both interpretations are acceptable, please mark the instance as N/A.

Sense inventory for *explain, subj*:

1. clarify, describe, make comprehensible
   e.g. *note, presentation, manual; people explaining things*

2. be a reason for something
   e.g. *action, fact; events explaining other events or states*

   Please mark as N/A all cases where either sense fits or it is unclear which sense is implied.

Sense inventory for *fall, subj*:

1. physically drop; move or extend downward
   e.g. *physical objects falling; also: extending downward, e.g. rainbow, light, hair*

2. decrease
   e.g. *price, inflation, profits, attendance*

3. lose power or suffer a defeat
   e.g. *Roman empire, Napoleon, France*

4. for a state (such as darkness or silence) to come, to commence
   e.g. *night, darkness, silence*

5. be categorized as or fall into a range
   e.g. *cases falling into a certain category, into several types, into a certain range*

6. be associated with or get assigned to a person or location or for event to fall onto a time
   - this sense is a metaphoric extension that covers all cases of one entity or event getting associated with another:
   e.g. *Birthdays, lunches, celebrations falling on a certain date or time.*
   e.g. *Stress or emphasis falling on a given topic or a syllable.*
   e.g. *Responsibility, luck, suspicion falling on or to a person.*

   Idiomatic expressions and phrasal verbs should be marked as N/A: e.g. fall from favor, fall through, fall in, fall back, fall silent, fall short, fall in love, etc.

Sense inventory for *fire, dobj*:

1. shoot, discharge a weapon
   e.g. *pistol, rifle*

2. shoot, propel a projectile
   e.g. *shot, bullet, rounds, spores*

3. dismiss from employment
   e.g. *firing people*

4. inspire
   e.g. *passion, imagination*

5. kick, hit, or pass an object of play in sports
   e.g. *rebound, goal, punch*

6. apply fire or fuel to; kindle
   e.g. *reactor, explosive, wood, clay*

   Sentences without direct object should be marked N/A: e.g. 'He was fired on'.

Sense inventory for *grasp, dobj*:

1. grab hold of something
   e.g. *arm, shoulders, barrel, sword*

2. understand, comprehend
   e.g. *significance, idea, intention*

3. seize an opportunity or chance
   e.g. *chance, offer, opportunity*

Sense inventory for *know, dobj*:

1. know the content
   e.g. *situation, answer*

2. be familiar or acquainted with something or someone
   e.g. *being acquainted with a person, familiar with a place or a feeling*

   Sentential arguments including infinitival closes (e.g. 'known to be') should be marked N/A. Other constructions that do not map to these senses (e.g. 'to be known for') should be marked N/A.

Sense inventory for *launch, dobj*:

1. physically propel into the air, water or space
   e.g. *missile, rocket*

2. begin or initiate an endeavor
   e.g. *campaign, inquiry, attack*

3. begin to produce or distribute a product; start a company
   - bring into existence a product or a company (extension of the 2nd sense)
   e.g. *release, edition, collection; newspaper, organisation*

Sense inventory for *lead, subj*:

1. for a person, to guide somebody to a destination by going with them
   e.g. *a leutentant leading his soldiers into battle, a host leading the guests into the living room*

2. for a person, to direct or preside over an activity or a group
   e.g. *project, group, protest, discussion*

3. cause or induce something; lead to a consequence
   e.g. *information leading to arrest, finding leading to a conclusion, impulse leading someone to do something*

4. for a path, to serve as a passage to
   e.g. *path, road, archway*

   Phrasals (e.g. 'lead up to') should be marked N/A.

Sense inventory for meet, *iobj_with*:

1. come together for a meeting with someone
   e.g. *manager, representatives, students*

2. encounter an event or experience a reaction, such as approval or dismay
   e.g. *success, resistance, interest, difficulty*

Appendix C

Test Data

The original, pre-annotation patterns for the annotated verbs are listed below. For each verb in the data set, one argument position was selected for analysis. We summarize the relevant senses for each verb, with some *selectional equivalents* and relevant collocates given for each sense. We give the mapping of each sense to several resources, including PropBank (Palmer et al., 2005), WordNet (Fellbaum, 1998), FrameNet (Fillmore et al., 2003; Ruppenhofer et al., 2006; Hiroaki, 2003), CPA patterns (Hanks and Pustejovsky, 2005; Rumshisky and Pustejovsky, 2006), and Sketch Engine word sketches (Kilgarriff et al., 2004) over the BNC. The entry for each sense includes the following:

(1) The sense gloss and the CPA-like pattern specification;

(2) The mapping of the sense to PropBank (PB), WordNet (WN), FrameNet (FN), OntoNotes (ON), and CPA patterns, if the corresponding sense is available in the resource;

(3) Argument sets manually identified using the Sketch Engine for the relevant argument position, from the BNC;

Remarks on lexical sets (coercion, semantic typing, etc.);

(4) Selectional equivalents for the sense;

(5) Translation equivalents.

C.1 Verbs

We list only the patterns relevant for a particular grammatical relation. The patterns (and the corresponding senses) which can be distinguished by virtue of syntactic

structure are omitted. Thus, for example, for *conclude*, a very dominant pattern, PERSON conclude that-CLAUSE is omitted, etc.

Note. The specification below assumes a simplified version of the CPA grammar as outlined in (Pustejovsky et al., 2004). Rather than using double brackets to indicate types, type names are capitalized. We used BULB (Havasi et al., 2006) to access WordNet and ProbBank. WordNet and PropBank entries in the sense summaries below are given in the order in which they are listed in BULB.

1. (*absorb*, `obj`)

   (1) PHYSOBJ | SUBSTANCE absorb SUBSTANCE
   Sense: absorb substance
   Resources: WN6 (become imbued), WN7 (take in), WN9 (suck or take up or in); CPA Pattern 1, 2 (absorbing nutrient or liquid); PB1 (suck up)
   Selectional equivalents: dissolve; sponge, soak up
   Translations: Rus. *rastvoriat', vpityvat'*

   `obj:` oil, oxygen, water, liquid, milk, carbon dioxide, acid, air, fluid, charcoal, soil, hydrogen sulphide, silica, salt, moisture, goodness, substance, antiserum, dirt, flavor; amount, percent, quantity

   (2) PHYSOBJ | SUBSTANCE absorb ENERGY
   Sense: absorb substance or energy
   Resources: WN9 (suck or take up or in); CPA Pattern 3, 4 (absorb energy or radiation)
   Selectional equivalents: emit, detect, transmit, focus, reflect, withstand, sense, measure; dissolve; sponge, soak up
   Translations: Rus. *pogloschat'*

   `obj:` radiation, heat, moonlight, sound, x-ray, wavelength, energy, impact, wave, shock, stress, flow, movement

   (3) ABSTRACT absorb RESOURCE
   Sense: consume a resource, such as time or money
   Resources: WN1; CPA Pattern 5
   Selectional equivalents: take up, consume
   Translations: Rus. *zanimat', pogloschat'*

   `subj:` operations, cuts, spending, expenditure, policies, insurance
   `obj:` percent, share, half, amount, pound, time

   (4) PERSON absorb ASSET
   Sense: bear the cost of; take on an expense
   Resources: WN8 (take up, as of debts or payments)

Selectional equivalents: assume, bear the cost of, compensate for, meet; offset, cover, recoup, recover
Translations: Rus. *pokryvat'*

`subj:` bidder, producer, member, consumer, investor, bank, HUMAN-GROUP
`obj:` sum, cost, loss, selling, tax, price increase, dollar, pound

(5) PERSON absorb ABSTRACT
Sense: learn skill or information
Resources: WN4 (take up mentally); CPA Pattern 7
Selectional equivalents: acquire, learn, assimilate; deduce, impart, digest, glean
Translations: Rus. *vbirat'*

`obj:` skill, information, mode, facts, rumours, culture

(6) ACTIVITY | TOPIC absorb PERSON
Sense: preoccupy (for a person to be preoccupied or immersed into something)
Resources: WN2 (consume all of one's attention or time); CPA Pattern 8
Selectional equivalents: occupy
Translations: Rus. *zanimat'*

`subj:` interest, plan, occupation, trimming, cutting `obj:` mind, thought, attention, PERSON

(7) REGION absorb PERSON
Sense: take in or assimilate, making part of a whole or a group
Resources: WN3 (assimilate or take in); CPA Pattern 6
Selectional equivalents: assimilate
Translations: Rus. *poglotit'*

`obj:` refugee, worker, employee, immigrant, PERSON

`BNC Frequency: dobj` 1213 / 2625

`Use:` Senseval-3 Subcategorization Acquisition

Comment:
WordNet sense with a generic gloss: "cause to become one with"

2. (*acquire*, `obj`)

(1) ENTITY | PERSON acquire ABSTRACT = QUALITY
Sense: take on a certain characteristic – form, attribute, or aspect
Resources: WN1 (take on a certain form, attribute, or aspect), WN2 (come to have or undergo a change of (physical features and attributes))

Selectional equivalents: take on, assume; gain, retain, possess, reveal, show
Translations: Rus. *priobresti, priniat'*

`obj:` facet, flavour, significance, quality, meaning, reputation, infection, patina, status, syndrome, taste, following, power, momentum, nickname, characteristic, stigma, pneumonia, importance, prominence, ability, character, experience

(2) ENTITY | PERSON acquire ABSTRACT = INFORMATION
Sense: learn
Resources: WN3 (gain knowledge or skills), WN4 (win something through one's efforts), WN5 (gain through experience)
Selectional equivalents: develop, gain, retain, possess, reveal, show
Translations: Rus. *priobresti*

`obj:` skill, knowledge, habit, expertise, competence, qualification, know-how, understanding, proficiency, competency, accent

(3) PERSON acquire PHYSOBJ = POSSESSION
Sense: purchase or become the ownwer of property; appropriate
Resources: WN6 (come into the possession of something concrete or abstract), PB1 (get, acquire)
Selectional equivalents: buy, purchase, own, transfer, lose, keep, increase, claim
Translations: Rus. *priobresti, kupit'; dostat', razdobyt'*

`obj:` asset, share, land, stake, property, information, rights, wealth, company, business, possession, subsidiary, estate, acre, weapon, stock

`BNC Frequency: dobj` 4126 / 6712

Mixed between senses 1 and 3:
`obj:` title, citizenship

Comment:
WordNet sense with a generic gloss: "come into the possession of something concrete or abstract"

3. (*admit*, `obj`)

(1) PERSON admit PROPOSITION
Sense: acknowledge the truth or reality of
Resources: PB1 (acknowledge truth), WN1 (declare to be true or admit the existence or reality or truth of)
Selectional equivalents: prove, accept, recognize, imply, deny, acknowledge, reveal, realize, confess, establish, conceal

Translations: Rus. priznat'
`obj:` defeat, assault, guilt, truth, ignorance, responsibility, mistake

(2) HumanGroup admit Person | PhysObj
Sense: grant entry to or accept into a community
Resources: PB2 (allow to enter), WN2 (admit into a group or community), WN3, WN4 (allow to enter; grant entry to), WN5, WN6
Selectional equivalents: accept, let in, allow in
Translations: Rus. *priniat'*
`obj:` student, patient, evidence, application

`BNC Frequency: dobj` 2369 / 10883

4. (*assume*, `obj`)

(1) Entity | Person assume Abstract = Quality
Sense: acquire a property or quality
Resources: WN2 (take on a certain form, attribute, or aspect), WN6 (occupy or take on a position), FN Adopt_selection (begin to use or take on some characteristic)
Selectional equivalents: take on, adopt, embrace; acquire, suggest, reveal, retain, reflect, indicate, gain, establish, determine
Translations: Rus. *priobresti, priniat'*

`obj:` significance, importance, proportion, status, shape, prominence, urgency, (alert, relaxed, sincere) expression, character, identity, guise, air, stance, position, pose

(2) Person assume Abstract = Responsibility
Sense: take on responsibility, position, or role
Resources: WN7 (seize or take control, take as one's right or possession, WN8 (take on as one's own the expenses or debts of another person), WN9 (take on titles, offices, duties, responsibilities)
Selectional equivalents: take on, take over, seize, usurp; win, secure, retain, inherit, attain, obtain, claim
Translations: Rus. *priniat', vziat' na sebya*

`obj:` mantle, presidency, control, power, chairmanship, leadership, title, office, command, post, throne, rights, kingship, rein; responsibility, burden, obligation, debt, duty, liability; function, role

Note: split responsibility set from presidency set

(3) Abstract | Person assume Proposition
Sense: implicitly incorporate or agree with a statement; take to be true

Resources: WN3 (take to be the case or to be true; accept without verification or proof), PB2 (believe)
Selectional equivalents: believe
Translations: Rus. *podrazumevat', imet' v vidu, predpolagat'*

`obj:` behaviorist position, knowledge, neutrality, rationality, connection, relationship, centrality, rate, validity, dichotomy

`BNC Frequency: dobj` 2748 / 10956

Mixed between senses:
`obj:` liability, position, neutrality

Comment:
The important distinctions here are between *assuming a significance* (which means, becoming significant) and *assuming a connection* (which means believing that there is a connection). These are clearly two different senses of *assume*. Assuming a role, an office, command, etc.

5. (*claim*, `obj`)

   (1) PERSON claim ABSTRACT = PROPOSITION
   Sense: assert the truth of propositional content
   Resources: WN2 (assert or affirm strongly; state to be true or existing), PB1 (assert)
   Selectional equivalents: assert, declare
   Translations: Rus. *utverjdat'*

   `obj:` originality, expertise, superiority, success, legitimacy, status, ownership, responsibility, dismissal, victory, right, priviledge, kinship, paternity

   (2) PERSON claim ENTITY = POSSESSION
   Sense: receive something
   Resources: WN5 (demand as being one's due or property; assert one's right or title to), WN4 (ask for legally or make a legal claim to, as of debts, for example), FN Claim_ownership (take or declare rightful ownership of), PB2 (seize)
   Selectional equivalents: receive, demand
   Translations: Rus. *poluchit', pretendovat' na +Acc.*

   `obj:` prize, reimbursement, suitcase, inheritance, share, money, attention, reward, payment, benefit, compensation, allowance, cost

   `BNC Frequency: dobj` 4345 / 18672

   Mixed between senses:
   `obj:` damages

Comment:
For some complements, such as *originality, expertise, superiority* the distinction between senses is clear, but there is a lot of boundary cases.

6. (*conclude*, `obj`)

   (1) PERSON | HUMANGROUP conclude EVENT | TIMEPERIOD
   Sense: finish
   Resources: PB2 (bring to an end), WN2 (bring to a close), causative of FN Process_end (come to an end), causative of WN5 (come to a close)
   Selectional equivalents: finish
   Translations: Rus. *zakonchit', zavershit'*

   `obj:` meeting, debate, negotiation, investigation, visit, tour, process, business, matter, conference, discussion; work, session, proceedings; chapter, section, article, novel, paper, letter, interview, speech, study, review; year, day, past

   (2) PERSON | HUMANGROUP conclude ABSTRACT = AGREEMENT
   Sense: enter into an agreement
   Resources: WN4 (reach an agreement on)
   Selectional equivalents: sign, agree on
   Translations: Rus. *zakluchit'*

   `obj:` treaty, agreement, deal, contract, truce, alliance, ceasefire, sale

   (3) PERSON conclude that-CLAUSE
   Sense: decide, establish something
   Resources: PB1 (decide), WN1 (decide by reasoning; draw or come to a conclusion), WN3 (reach a conclusion after a discussion or deliberation), FN Coming_to_believe (arrive at a judgement or opinion by reasoning)
   Selectional equivalents: decide
   Translations: Rus. *reshit'*

   `BNC Frequency: dobj` 815 / 5552 (including clausal arguments)

   Comment:
   The last sense does not usually allow NP-complements, except in very rare cases, as in: "Nor would it be justifiable to conclude some causal connection between smoking and lung cancer on the evidence of just one heavy smoker contracting the disease."

7. (*cut*, `obj`)

   (1) PERSON | HUMANGROUP cut ENTITY = QUANTIFIABLE
   Sense: reduce

Resources: PB2 (reduce), WN8 (cut down on; make a reduction in)
Selectional equivalents: reduce, bring down
Translations: Rus. *urezat', umen'shit'*

`obj:` cost, emission, spending, price, rate, deficit, budget, workforce, crap, tax, subsidy, expenditure, jobs, losses, pay, consumption, wages, overhead, production, polution, tariff, grant, bill, fee, odds, output, funding, intake, staff

(2) PERSON | PHYSOBJ cut PHYSOBJ
Sense: physically cut, separate pieces of or make an incision
Resources: PB1 (slice), WN20 (separate with or as if with an instrument), WN22 (make an incision or separation), FN Cause_harm (make an opening, incision, or wound in (something) with a sharp tool or object)
Selectional equivalents: touch
Translations: Rus. *razrezat', porezat'*

`obj:` throat, hair, grass, ice, corner (metaphoric), cake, nail, ribbon, lawn, meat stem, wire, rope, wood, bread, cord, hedge, tape, pipe, orange, tomato, tree, finger, hay, artery, wrist, flower, tile

(3) PERSON | HUMANGROUP cut PHYSOBJ
Sense: cut out a shape or a form (result of cutting)
Resources: WN28 (form or shape by cutting or incising)
Selectional equivalents: tear, rip
Translations: Rus. *vyrezat', prorezat'*

`obj:` swathe, hole, slit, groove, slot; slice, strip, piece, length, cross-section; portion

`BNC Frequency: dobj` 7099 / 17863 (including phrasals)

Idiomatic, infrequent, or metaphoric uses:
obj: queue, tooth, class, engine, noise

Phrasal verbs, idioms, or syntactic patterns to be separated out:
cut off, *cut short* (story, visit, honeymoon), *cut loose*, *cut from* SCENE *to* SCENE, *cut through*, *cut* PERSON *from*, etc.

8. (*deny*, `obj`)

(1) PERSON 1 deny TOPTYPE to PERSON 2 | HUMANGROUP
Sense: refuse to give or grant something
Alternate pattern: PERSON 1 deny {PERSON 2 | HUMANGROUP} TOPTYPE
Resources: WN1 (refuse to grant, as of a petition or request), WN2 (deny

oneself (something); restrain, especially from indulging in some pleasure: abnegate), WN3 (refuse to let have: refuse), PB1 (turn down, reject)
Selectional equivalents: refuse, grant, approve
Translations: Rus. *otkazat' v +Prep*

`obj:` access, aid, allowance, approval, asylum, bail, consent, council, credit, entry, exemption, extension, fund, funding, honour, information, interview, license, opportunity, option, permission, recognition, registration, request, visa, vote

(2) PERSON deny PROPOSITION
Sense: state or maintain that something is untrue
Alternate pattern: PERSON deny that-CLAUSE
Resources: FN Statement (claim that something is false; LUs:*acknowledge.v, admit.v, affirm.v, assert.v, confirm.v, maintain.v, proclaim.v, reaffirm.v, reiterate.v*), WN6 (refuse to accept or believe), WN7 (declare untrue; contradict)
Selectional equivalents: refute, confirm, admit, assert, negate, affirm, verify
Translations: Rus. *otricat' +Acc*

`obj:` accusation, assault, attack, depression, difference, distinction, effectiveness, effects, error, fact, figures, findings, guilt, hypothesis, impact, importance, improvement, inference, influence, intention, intent, interest, interpretation, involvement, kidnapping, killing, link, matter, negligence, potential, prejudice, presence, problem, reality, report, rumour, seriousness, severity, significance, speculation, statement, story, suggestion, tension, theory, trend, truth, value, weakness, word

(3) PERSON deny ABSTRACT
Sense: refuse to acknowledge or follow something, reject
Resources: WN5 (refuse to recognize or acknowledge)
Selectional equivalents: reject; renounce; elude, escape; denounce
Translations: Rus. *otvergat' +Acc; otrech'sya ot +Gen; otkazat'sya ot +Gen*

`obj:` classification, destiny, ideal, belief, conviction, culture, faith, Jesus, Lord, Devil, body, love

`BNC Frequency: dobj` 3611 / 7509

Comment:
Many nouns allow multiple interpretations, and additional context is required for disambiguation, e.g.:

We were denied a view of the ocean (Sense 1)
He would not deny his own traditional view (Sense 2)

Hospitals may deny help to older people (Sense 1)
He would not deny the help of a physician (Sense 3)

The government denies equal conditions to workers (Sense 1)
They deny the actual conditions of black oppression (Sense 2)

Asylum-seekers might be denied refugee status (Sense 1)
Mental disorders that explain crimes deny their very status as crimes (Sense 2)

As is often the case with verbs, the distinction between the senses is sometimes blended:

deny primacy to altruism
deny the primacy of altruism

No one will deny this quality to him
No one will deny this quality of his

deny validity to this system
deny the validity of this system

Ditransitive construction for Sense 1 is dominated by the use in passive:
e.g. *Our brothers are denied freedom of worship*

9. (*dictate*, `obj`)

   (1) PERSON | HUMANGROUP dictate INFO · PHYSOBJ
   Sense: verbalize to be recorded
   Resources: WN1 (say outloud for the purpose of recording)
   Selectional equivalents: read
   Translations: Rus. *diktovat'*

   `obj:` passage, story, letter, memoirs, novel, message, text, note, words, account, work

   (2) HUMANGROUP | ABSTRACT dictate ABSTRACT
   Sense: determine the character of or serve as motivation for
   Resources: PB1 (to impose or command), WN2 (issue commands or orders for)

Selectional equivalents: determine, control, suggest, motivate, influence, specify
Translations: Rus. *opredeliat', diktovat'*

`obj:` game, number, rate, level, position, life, way, pace, term, choice, event, hour, policy, shape, caution, action, treatment, tactic, pattern, property, price, course, curriculum, move, design, method, quality, approach, arrangement, use, structure, decision, behavior; nature, kind, extent

`subj:` music, question, convention, economics, subject matter, consideration, price, circumstances, security, custom, prudence, tradition, conscience, common sense, logic, wisdom, fashion

`BNC Frequency: dobj` 477 / 1264

Comment:
The first sense has PERSON in Dative / Accusative alternation:
e.g. *The teachers aren't allowed to dictate the children.*

10. (*drive*, `obj`)

(1) PERSON drive VEHICLE
Sense: direct a vehicle's motion by its controls
Resources: WN15 (operate or control a vehicle), causative of WN18 (move by being propelled by a force), PB1 (drive a vehicle: vehicle or path), causative of FN Self_motion (move under its own power or directed by a driver), MW4b (to operate the mechanism and controls and direct the course of (as a vehicle))
Selectional equivalents: park, reverse, stop
Translations: Rus. *vesti, vodit'*

`obj:` car, vehicle, minibus, truck, van, tractor, jeep, volvo, porsche, ambulance, chariot, cab, cart, bus, taxi, wagon, train, bike, crane, tank, diesel, horse

(2) PERSON drive DISTANCE
Sense: travel in a vehicle a certain distance
Resources: PB1 (drive a vehicle: vehicle or path)
Selectional equivalents: travel, walk, cover
Translations: Rus. *proexat'*

`obj:` mile, yard, kilometer, way, distance

(3) PERSON drive PERSON | PHYSOBJ (ADV[LOCATION])
Sense: give a lift, transport

Resources: WN13 (travel or be transported in a vehicle), WN17 (cause someone or something to move by driving), FN Bringing (convey in a car), MW4c (to convey in a vehicle)
Selectional equivalents: transport
Translations: Rus. *vezti, vozit', podvezti, otvezti*

`obj:` bomb, passenger, PERSON,

(4) MECHANISM | POWER drive MECHANISM
Sense: provide power for and/or physically move
Resources: WN4 (cause to function by supplying the force or power for or by controlling), MW3c (to set or keep in motion or operation): drive machinery by electricity
Selectional equivalents: power, control, activate, move
Translations: Rus. (different translations, depending on the mechanism driven) *upravliat'*

`subj:` wheel, (coal, electric) power, wind, cowling, jet, motor; steam, vapour, gas, energy, motor, petrol, waterwheel, (gas, propulsion, front, two-stroke) engine

`obj:` propeller, machinery, wheel, generator, waterwheel, turbine, motor, shaft, engine, mill, rotor, pump, load, carriage

(5) POWER drive {PHYSOBJ = VESSEL} ADV[DIRECTION]
Sense: force a vessel to move in a direction
Resources: MW3a (to impart a forward motion to by physical force: 'waves drove the boat ashore')
Selectional equivalents: carry, bring
Translations: Rus. *gnat'*

`obj:` tanker, boat, vessel, ship, clouds, sea

`subj:` wind, gale, storm, tide

(6) PERSON drive {ANIMATE = ENEMY} ADV[DIRECTION]
Sense: force adversary to leave
Resources: WN8 (cause to move back by force or influence)
Selectional equivalents: force, face
Translations: Rus. *vygnat'*

`obj:` man, dog, enemy, russians, romans, army, Iraq, french, english, competitor, inhabitants

(7) PERSON drive {ANIMATE = CATTLE | GAME} ADV[DIRECTION]
Sense: physically force animal to go
Resources: WN16 (urge forward)
Selectional equivalents: chase, herd
Translations: Rus. *gnat', peregonyat'*

`obj:` team (of horses), cattle, sheep, flock (of sheep), deer, donkey, animals, herd, cow, beast, elephant

(8) {ANIMATE | ABSTRACT = CAUSE} drive {ANIMATE | ABSTRACT} {ADV[STATE] | to-INF}
Sense: cause or force into a state or activity
Resources: WN11 (force into or from an action or state, either physically or metaphorically), WN12 (compel somebody to do something, often against his own will or judgment), MW5b (to compel to undergo or suffer a change (as in situation or emotional state: 'drove him crazy') MW5d (to press or force into an activity, course, or direction)
Selectional equivalents: make
Translations: Rus. *vvesti v +Acc, vynudit' +*INF*, zastavit' +*INF*, zastavit'*
`obj:` men, peasants, people, male; price, inflation, thought
`subj:` curiosity, greed, arrogance, necessity, recession, policies, PERSON
`adv:` into despair, into poverty, into debt, off the land, wild, mad, together, away, to-CLAUSE (to climb corporate ladders, to face dangers); from her mind, up, down, high, through the roof, to equality

(9) PERSON | PHYSOBJ drive PHYSOBJ = NAIL into PHYSOBJ
Sense: push a sharp object into another object
Resources: WN10 (push, propel, or press with force)
Selectional equivalents: hammer, put into
Translations: Rus. *zabit', vstavit'*
`obj:` nail, wedge, spike, stake, screw, peg, sword, needle, fist

(10) PERSON drive PHYSOBJ ADV[DIRECTION]
Sense: strike or throw an object of play
Resources: WN9 (cause to move rapidly by striking or throwing with force)
Selectional equivalents: deliver, miss
Translations: Rus. *zabit'*
`obj:` ball, shot, kick

`BNC Frequency: dobj` 5733 / 14796

Comment:

Diachronically motivated boundary case:
Drive a horse (drive cattle vs. drive a vehicle)

Combining semantics of different arguments disambiguate the predicate:
Engine | PERSON *drives a car* ADV

Mixed between senses (poor disambiguators):
mother, dog, man, victim, boy, girl, wife, lover:

Drive the men home | to a meeting
Drive the men up the hill
Drive the men into despair

True ambiguity, resolved only in extended context:
He drove the men up the hill (transport vs. force an adversary to leave)

Infrequent:
drive a tunnel (= bore)

11. (*edit*, `obj`)

    (1) PERSON edit DOCUMENT
    Sense: make changes to the text; modify content; make corrections
    Resources: PB1 (edit, work on text, etc), WN3 (cut and assemble the components of), CPA Pattern 1, 2 (make changes to the text of document)
    Selectional equivalents: write, copy
    Translations: Rus. *redaktirovat'*

    `obj:` note, document, letter, work, text, file, statement, data, reply, script, writing, contribution, program; tape, video, talk, commercial, picture, screen, shot

    (2) PERSON edit BOOK | PERIODIC PUBLICATION
    Sense: supervise publication
    Resources: WN4 (supervise the publication of), WN1 (prepare for publication or presentation by correcting, revising, or adapting), CPA Pattern 3, 4 (person is responsible for preparing the content for publication)
    Selectional equivalents: publish, produce
    Translations: Rus. *redaktirovat'*

    `obj:` journal, newspaper, magazine, program, page, column, newsletter, series; book, volume, edition, study, guide, letters, compendium, selection, proceedings, collection

    `BNC Frequency: dobj` 720 / 1661

12. (*enjoy*, `obj`)

    (1) ANIMATE enjoy EVENT
    Sense: to like something
    Resources: WN1 (take delight in), WN3 (derive or receive pleasure from; get enjoyment from; take pleasure in), WN4 (get pleasure from), ON1 (relish, savor, delight in), PB1 (take pleasure from), FN Experiencer_subj (take pleasure in)

Selectional equivalents: like, love
Translations: Rus. *nravit'sya +Nom*

`obj:` taste, food, dancing, outdoors, barbecues, thrill, countryside, scenery, spectacle, sensation, school, practice, trip, holiday, visit

(2) ANIMATE | ABSTRACT enjoy ABSTRACT = STATE
Sense: to have something
Resources: WN2 (have benefit from), WN5 (have for one's benefit), ON2 (derive benefit from)
Selectional equivalents: guarantee, demand; gain, achieve, maintain, earn
Translations: Rus. *dobit'sya +Acc*

`obj:` protection, health, security, monopoly, comfort, lifestyle, notoriety, acclaim, fame, independence, autonomy, liberty, prosperity, wealth, stability

`BNC Frequency: dobj` 9915 / 14212

Mixed between senses 1 and 2:
`obj:` freedom, atmosphere

`Use:` Semeval-2007 (English Lexical Sample, English SRL, English All-Words Tasks)

13. (*explain*, `subj`)

(1) PERSON | INFO | ABSTRACT = THEORY explain EVENT
Sense: clarify, describe, account for
Resources: WN2 (make plain and comprehensible), ON1 (clarify, make comprehensible, describe), FN Statement (make (something) clear by describing it in more detail)
Selectional equivalents: suggest, describe, predict, specify
Translations: Rus. *ob'yasnyat'sya +Prep, ob'yasnyat' +Nom*

`subj:` leaflet, booklet, manual, guide, note, letter, handbook, chapter, pamphlet, paragraph, section, article, brochure, legend, introduction, dictionary, statement, notice, sentence, report; theory, hypothesis, metaphor, example, approach, principle, analysis, appeal, study, idea, science; spokesman, solicitor, doctor, sergeant, colonel, minister, teacher, commentator, angel, nurse, author, mother, trainer, tutor, officer, assistant, father, PERSON; voice

(2) EVENT | ABSTRACT explain EVENT
Sense: be a reason for, account for
Resources: WN1 (serve as a reason or cause or justification of), ON2 (be

a reason, cause, or justification of), FN Explaining_the_facts (to make the existence of a state of affairs plain or understandable)
Selectional equivalents: account for, justify
Translations: Rus. *ob'yasnyat'*

`subj:` reference, fact, variation, difference, combination, presence, infection, process, fall, complexity, absence, characteristic, tendency, phenomenon, pollution, illness, difficulty, intention, increase, disease

`BNC Frequency: subj` 13027 / 18664

Comment:
Sense 2 is often used in passive, with *by*

Mixed between senses:
reason, account

`Use:` Semeval-2007 (English Lexical Sample, English SRL, English All-Words Tasks)

14. (*fall*, `subj`)

(1) PhysObj fall (adv[direction])
Sense: physically drop; move or extend downward
Resources: PB1 (move downward), WN12 (move downward and lower, but not necessarily all the way), WN13 (descend in free fall under the influence of gravity), WN32 (fall from clouds), WN28 (touch or seem as if touching visually or audibly), FN Motion_directional (move from a higher to a lower level, typically rapidly and without control)
Selectional equivalents: drop, hit
Translations: Rus. *padat', upast'*

`subj:` rain, snow, bomb, hair, burden (metaphorical), axe, tear, drizzle, leaf, curtain, roof, meteorite, blossom, crumb, debris, trousers, skirt, dress, cloak, hat, snowflake, sky, particle, flake, tree, drop, rider, screwdriver, petal, stone, glass, hand, head, blanket, hammer, wall, lid, horse, ceiling, rock, casualty; blow; shadow, light, sunlight, radiation, sun; eye, gaze, glance

(2) Abstract = Measurable fall
Sense: decrease
Resources: WN2 (decrease in size, extent, or range), FN Change_position_on_a_scale (decrease)
Selectional equivalents: decrease, diminish, lessen, increase, rise, decline, climb, dip, gain, skyrocket, reach, triple, double, fluctuate
Translations: Rus. *ponizit'sya, upast'*

`subj:` turnover, share, price, profit, output, temperature, inflation, unemployment, index, rate, sales, wages, income, vote, stock market, import, surplus, yield, ratio, tax, cost, circulation, membership, demand, spending, earnings, attendance, rating, profitability, popularity; standard, level

(3) ABSTRACT fall
Sense: for a state, to come, begin, commence
Resources: WN3 (come as if by falling)
Selectional equivalents: begin, come, go
Translations: Rus. *nastupit'*

`subj:` dusk, darkness, hush, silence, night

(4) EVENT | ABSTRACT fall on TIME | LOCATION
Sense: occur at a specified time or location
Resources: PB2 (occur: when/where), WN5 (occur at a specified time or place)
Selectional equivalents: occur, precede
Translations: Rus. *vypast', popast', prixodit'sya*

`subj:` birthday, cancellation, stress, emphasis

(5) EVENT | ENTITY = CHOICE fall on | to PERSON | ENTITY = CATEGORY
Sense: be assigned to someone or associated with a category
Resources:
Selectional equivalents:
Translations: Rus. *dostat'sya +Dat, vypast' +Dat, vypast' na +Dat, popast' v +Acc*

`subj:` work, burden, choice; research

(6) PERSON | HUMANGROUP fall
Sense: lose power or suffer a defeat
Resources: PB5 (be defeated)
Selectional equivalents: be defeated
Translations: Rus. *past'*

`subj:` Hussein, France, empire

`BNC Frequency: subj` 22318 / 26296

Comment:
An adverbial or a prepositional phrase may change the meaning completely for certain types of subjects, e.g. PERSON:
fall into a trap, from grace, from the roof, etc.

Phrasal verbs, idioms, or syntactic patterns to be separated out:
fall through, *fall in*, *fall back*, *fall silent*, *fall short*, *fall in love*, etc.

Idiomatic, infrequent, or metaphoric uses:
`subj:` suspicion, responsibility; face, plea

15. (*fire*, `obj`)

    (1) PERSON fire FIREARM
    Sense: shoot, discharge a weapon
    Resources: WN3 (cause to go off), FN Use_firearm (discharge (a gun or other weapon)), CPA Pattern 1 (person causes firearm to discharge projectile toward target)
    Selectional equivalents: shoot
    Translations: Rus. *vystrelit' iz +Gen*

    `obj:` gun, revolver, rifle, pistol, bow, barrel, blaster, mortar, weapon, cannon, shotgun

    (2) PERSON | FIREARM fire PHYSOBJ = PROJECTILE
    Sense: shoot, propel a projectile
    Resources: PB1 (fire a gun), WN3 (cause to go off) FN Shoot_projectiles (propel (a bullet or projectile) from a gun or other weapon), CPA Pattern 2 (person causes firearm to discharge projectile toward target)
    Selectional equivalents: shoot, throw
    Translations: Rus. *zapustit'*

    `obj:` shot, round, bullet, grenade, flare, blast, burst, spray, stream, ball, torpedo, rocket, missile, blank, shell, cartridge, charge; barrage, volley; staple, pin, nail; smile

    (3) PERSON 1 fire PERSON 2
    Sense: dismiss from employment
    Resources: PB3 (cause to cease employment), WN9 (terminate the employment of), FN Firing (dismiss from a job), CPA Pattern 6 (dismiss from employment)
    Selectional equivalents: dismiss, lay off, sack, terminate
    Translations: Rus. *uvolit'*

    `obj:` people, staff, colleague, employee, worker, official, person, lieutenant-colonel, PERSONAL-PRONOUN

    (4) TOPTYPE fire PERSON's ENTHUSIASM
    Sense: inspire
    Resources: CPA Pattern 12 (person is filled with enthusiasm because of something)
    Selectional equivalents: inspire
    Translations: Rus. *zazhech'*

    `obj:` enthusiasm, imagination, interest, sense, heart, motivation

(5) PERSON fire BALL (ADV[DIRECTION])
Sense: kick, hit, or pass (in sports)
Resources: CPA Pattern 14 (kick, hit, or pass the ball in a specific direction)
Selectional equivalents:
Translations: Rus. different translations
`obj:` ball, winner, rebound, cross, goal

(6) PERSON fire PHYSOBJ
Sense: apply fire or fuel to; kindle
Resources: WN (bake in a kiln so as to harden)
Selectional equivalents: heat, burn
Translations: Rus. *podzhech', obzhech'*
`obj:` reactor, explosive, wood, clay

`BNC Frequency: dobj` 1124 / 3360

16. (*grasp*, `obj`)

(1) PERSON grasp PHYSOBJ
Sense: take hold of
Resources: CPA Pattern 1, 2 (seize hold of something), WN2 (hold firmly), FN Manipulation (seize and hold firmly)
Selectional equivalents: grab, grip, clutch
Translations: Rus. *sxvatit'*
`obj:` horns, edge, device, handle, chain, barrel, side, suitcase, arm, knife, saddle, bottle, string, shield, spear; collar, coat; face, hand, leg, knee, shoulder; messenger, coroner, soldier, PERSONAL-PRONOUN, PERSON

(2) PERSON grasp IDEA
Sense: understand
Resources: CPA Pattern 4, WN1 (get the meaning of something), FN Grasp (comprehend fully)
Selectional equivalents: comprehend, understand
Translations: Rus. *ponyat'*
`obj:` know-how, size, nature, question, potential, possibility, structure, difference, point, fact, distinction, idea, principle, vulnerability, reality, complexity, meaning, viewpoint, truth, definition, pun, association, implication, situation, religion, enormity, extent, concept, relation, relationship, interconnections, prospect, dimension, vision, picture, importance, strategy, basics, issue, interplay, magnitude, form, thesis, essentials, ramifications, view, model, significance

(3) PERSON grasp OPPORTUNITY
Sense: seize an opportunity or chance
Resources: CPA Pattern 6
Selectional equivalents:
Translations: Rus. *ne upustit' +Acc*
`obj:` opportunity, chance, offer, moment

`BNC Frequency: dobj` 1112 / 1613

17. (*know*, `obj`)

(1) PERSON know ABSTRACT = PROPOSITION
Sense: be aware of something; knowledge of content
Resources: FN Awareness (be aware of through observation, inquiry, or information), PB1 (understand), WN3 (be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about), WN4 (be aware of the truth of something; have a belief or faith in something; regard as true beyond any doubt), WN6 (have fixed in the mind)
Selectional equivalents: believe
Translations: Rus. *znat'*
`obj:` answer, name, story, word, truth, meaning, phenomenon, identity, secret, thing, fact, way, difference, rule, technique, process, situation, view, detail, outcome, result, value, reason, cause, score, basics, address, odds, date, limitations, circumstances, area, destination, contents

(2) PERSON know PERSON | ENTITY
Sense: be familiar with something or someone; familiarity, acquaintance
Resources: FN Familiarity, WN2 (be familiar or acquainted with a person or an object), WN7 (have firsthand knowledge of states, situations, emotions, or sensations), WN8 (perceive as familiar)
Selectional equivalents: remember
Translations: Rus. *znat'*
`obj:` people, father, man, girl, mother, lady, woman, person, mum, bloke, guy, family, enemy, author, artist, PERSON; feeling, happiness, love; area, place, LOCATION;

`BNC Frequency: dobj` 33732 / 178296 (including clausal arguments)

Mixed between senses:
A lot of borderline cases between knowing the content of something and knowing of the existence of something (familiarity):
`obj:` trick, technique, position, theory, poem

Comment:
Coercions from PERSON are poor disambiguators, e.g. *Do you know Chomsky?* (person or theory).

Adverbs contribute to disambiguation: awareness of content: *exactly, for sure, precisely, instinctively*

Idiomatic, infrequent or metaphoric: *know one's place*

18. (*launch*, `obj`)

    (1) PERSON launch PHYSOBJ (ADV[LOCATION])
    Sense: physically propel into the air, water or space
    Resources: WN3 (propel with force), FN Shoot_projectiles (end out or hurl forcefully)
    Selectional equivalents: propel, dispatch
    Translations: Rus. *zapustit'*

    `obj:` satellite, rocket, missile, lifeboat, Sputnik, torpedo, boat, shuttle, bomb, carrier, craft, yacht, ship

    (2) PERSON launch EVENT
    Sense: begin or initiate an endeavor
    Resources: WN4 (get going; give impetus to)
    Selectional equivalents: initiate, begin; organize, mastermind, spearhead, orchestrate, mount; commence, initiate, instigate, intensify, complete, undertake
    Translations: Rus. *nachat', initsiirovat'*

    `obj:` expedition, campaign, initiative, project, investigation, drive, competition, exhibition, quest, effort, phase, defense; attack, assault, offensive, raid, invasion, rebellion, crusade, witch-hunt, war, protest, revolution; inquiry, appeal, bid, petition, review, action

    (3) PERSON launch PRODUCT | COMPANY
    Sense: begin to produce or distribute (a subcase of launch EVENT); found a company
    Resources: PB1 (introduce, bring up, start), WN6 (set up or found)
    Selectional equivalents: introduce, release
    Translations: Rus. *osnovat', vypustit', zapustit' v prodazhu*

    `obj:` magazine, perfume, release, edition, workstation, system, car, venture, product, club; manifesto, publication, newspaper, journal, company; festival

`BNC Frequency: dobj` 3474 / 6633

Comment:
Sense 3 may be considered a conventionalized coercion, where PRODUCT is coerced to a PRODUCT-launching EVENT: the distinction for many cases is not very clear.

Additional indicators of sense 1:
PPs *from, toward*

Mixed between senses 1 and 2:
`obj:` ship (expedition or physical object)

Example of empty headwords as poor disambiguators:
`obj:` *series of monographs* vs. *series of attacks*

19. (*lead*, `subj`)

(1) PERSON 1 lead PERSON 2 | HUMANGROUP ADV[DIRECTION]
Sense: guide somebody to a destination by going with them
Resources: PB1 (directed motion: cause to go), WN6 (take somebody somewhere: conduct, direct, guide, lead, take), FN Cotheme (show (someone) the way to a destination by preceding or accompanying them; LUs: *accompany.v, chase.v, conduct.v, escort.v, follow.v, guide.v, pursue.v, walk.v*)
Selectional equivalents: accompany, follow, escort, walk
Translations: Rus. *vesti, otvesti, provesti, preprovodit'*
`subj:` PERSON

(2) PERSON 1 lead EVENT | HUMANGROUP
Sense: direct or preside over an activity
Resources: PB2 (act as a project leader), WN2 (preside over: chair, lead, moderate), WN7 (be in charge of: head, lead), WN3 (lead as in the performance of a composition (conduct, direct, lead), FN Leadership (be in charge or command of; LUs: *command.v, govern.v, head.v, preside.v, reign.v, rule.v, run.v*)
Selectional equivalents: run, head, preside over, command
Translations: Rus. *rukovodit' +Inst*
`subj:` PERSON

(3) EVENT | ABSTRACT = CAUSE lead to EVENT = RESULT
Sense: cause or induce something; lead to a consequence
Alternate pattern: EVENT | ABSTRACT = CAUSE lead PERSON to INF | into EVENT | STATE
Resources: PB3 (resulted), WN1 (cause to undertake a certain action), WN8 (be conducive to), WN9 (have as a result or residue), WN10 (tend to or result in), FN Causation ("One thing leads to another"; a rather

vaguely defined sort of causation, although often used in quite definite cases: "Smoking leads to higher rates of lung cancer"; LUs: *bring on.v, bring.v, bring_about.v, cause.v, induce.v, precipitate.v*)
Selectional equivalents: induce, cause
Translations: Rus. *privesti k +Dat*

`subj:` event, factor, circumstance, condition, attitude, experience, situation, fact, behavior, commercialism, puritanism, reasoning, process, approach, course, research, reform, policy, study, proposal, economy, action, change, operation, regulation, development, information, consideration, observation, experiment, finding, negotiation, discussion, argument, investigation, exercise, talk, incident, scandal, blunder, conduct, volatility, heredity, evidence; impulse, instinct

(4) PATH lead to LOCATION | ADV[LOCATION]
Sense: Serve as a passage to
Resources: WN11 (stretch out over a distance, space, time, or scope; run or extend between two points or beyond a certain point: extend, go, pass, run), WN12 (lead, extend or afford access)
Selectional equivalents: go, run
Translations: Rus. *vesti, prostirat'sya*

`subj:` path, road, track, lane, street, staircase, stairway, corridor, tunnel, footpath, passage, trail, alley, ramp, archway, turn, pathway, driveway, ladder, avenue, doorway, door, gate, steps, backstreet, slope, hill, ridge

`BNC Frequency: subj` 26100 / 32658

20. (*meet*, `iobj_with`)

(1) PERSON meet with PERSON
Sense: conduct a meeting
Resources: WN8 (get together socially or for a specific purpose), WN5 (come together), PB3 (get together (with)), FN Congregating (to come into the company of, come together with)
Selectional equivalents: see
Translations: Rus. *vstretit'sya*

`pp_with-p:` leader, representative, officials, delegation, counterpart, minister, president, secretary, company, staff, department, union, people, head, manager, Arafat, Bush, Gorbachev

(2) PERSON | ABSTRACT meet with ABSTRACT = REACTION | EVENT
Sense: encounter an event or experience a reaction
Resources: WN13 (experience a reaction)

Selectional equivalents: encounter, face
Translations: Rus. *byt' vstrechennym s +Inst*; varied translations

`pp_with-p:` success, approval, opposition, resistance, hostility, incredulity, silence, derision, acclaim, refusal, enthusiasm, obstruction, scepticism, laughter, shrug, criticism, dismay, protest, excuse, demands, interest; stare, look (ADJ) response, reaction; degree, deal; setback, accident, failure, problem, difficulty

`BNC Frequency: iobj_with` 1395 / 32929

Comment:
The sense corresponding to WN10 (get to know; get acquainted with), PB2 (kennenlernen) occurs only with direct objects.

Bibliography

E. Agirre and P. Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Text, Speech and Language Technology. Springer Netherlands, July.

E. Agirre and A. Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic, June. Association for Computational Linguistics.

E. Agirre, D. Martínez, O. López de Lacalle, and A. Soroa. 2006. Two graph-based algorithms for state-of-the-art wsd. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 585–593, Sydney, Australia, July. Association for Computational Linguistics.

E. Agirre, L. Màrquez, and R. Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, June.

E. Amigó, J. Gonzalo, and J. Artiles. 2008. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Techical report, Departamento de Lenguajes y Sistemas Informticos (UNED), Madrid, Spain.

Yu. Apresjan, I. Mel'chuk, and A. Zholkovsky. 1969. Semantics and lexicography: Towards a new type of unilingual dictionary. *Studies in Syntax and Semantics*, pages 1–33.

Ju. Apresjan. 1973. Regular polysemy. *Linguistics*, 142(5):5–32.

Ju. Apresjan. 1974. *Leksicheskaja Semantika. Sinonimicheskie sredstva jazyka*. Nauka, Moskva.

Ju. Apresjan. 2000. *Systematic Lexicography*. Oxford University Press.

BNC. 2000. *The British National Corpus*. The BNC Consortium, University of Oxford, *http://www.natcorp.ox.ac.uk/*.

T. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, May 2002*, pages 1499–1504.

F. Busa, N. Calzolari, and A. Lenci. 2001. Generative Lexicon and the SIMPLE Model: Developing Semantic Resources for NLP. Oxford University Press, Cambridge.

X. Carreras and L. Marquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL-2004*, pages 89–97.

X. Carreras and L. Marquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-2005*.

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

T. Cover and J. Thomas. 1991. *Elements of Information Theory*. John Wiley & sons.

D. A. Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Patrick St. Dizier and Evelyne Viegas, editors, *Computational Lexical Semantics*, pages 33–49. Cambridge University Press, Cambridge, England.

D. A. Cruse. 2000. *Meaning in Language, an Introduction to Semantics and Pragmatics*. Oxford University Press, Oxford, United Kingdom.

J. Curran. 2004. *From Distributional to Semantic Similarity*. PhD dissertation.

I. Dagan. 2000. Contextual word similarity. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*. Marcel Dekker.

B. Dorow and D. Widdows. 2003. Discovering corpus-specific word-senses. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages Conference Companion pp. 79–82, Budapest, Hungary, April.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

C. Fellbaum, editor. 1998. *Wordnet: an electronic lexical database*. MIT Press.

C. Fillmore, C. Johnson, and M. Petruck. 2003. Background to Framenet. *International Journal of Lexicography*, 16(3):235–250, September.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Oxford: Philological Society.

T. Fontenelle. 1998. Discovering significant lexical functions in dictionary entries. In A. P. Cowie, editor, *Phraseology. Theory, Analysis and Applications*, pages 189–207. Oxford: Clarendon Press.

W. Gale, K. Church, and D. Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

P. Gamallo, A. Agustini, and G. Lopes. 2005. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–145.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, Norwell, MA, USA.

M. Halliday. 1973. *Explorations in the Functions of Language.* London: Edward Arnold.

P. Hanks and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée.*

P. Hanks. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics*, 1(1).

Z. Harris. 1985. Distributional structure. In J. Katz, editor, *Philosophy of Linguistics*, pages 26–47. Oxford University Press, New York.

T. Hastie, R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning.* Springer-Verlag, New York, NY.

C. Havasi, J. Pustejovsky, and M. Verhagen. 2006. BULB: A unified lexical browser. In *LREC 2006, Genoa, Italy.*

D. Hindle. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, PA. Association for Computational Linguistics.

S. Hiroaki. 2003. FrameSQL: A software tool for FrameNet. In *Proceedigns of ASIALEX '03*, pages 251–258, Tokyo, Japan. Asian Association of Lexicography.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

A. Kilgarriff and D. Tugwell. 2001. Word sketch: Extraction and display of significant collocations for lexicography.

A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex, Lorient, France*, pages 105–116.

A. Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31:91–113.

S. Landes, C. Leacock, and R.I. Tengi. 1998. Building semantic concordances. In C. Fellbaum, editor, *Wordnet: an electronic lexical database*. MIT Press, Cambridge (Mass.).

A. Lascarides, A. Copestake, and T. Briscoe. 1996. Ambiguity and Coherence. *Journal of Semantics*, 13(1):41–65.

LDOCE. 1978. Longman Dictionary of Contemporary English. Longman Group Ltd, Harlot, England.

L. Lee. 1999. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, et al. 2000. SIMPLE: A GENERAL FRAMEWORK FOR THE DEVELOPMENT OF MULTILINGUAL LEXICONS. *International Journal of Lexicography*, 13(4):249–263.

D. Lin. 1998. Automatic retrieval and clustering of similar words. *COLING-ACL, Montreal, Canada*.

D. Lin. 2002. Dependency-based thesaurus. Dekang Lin's Home Page. Retrieved August 3, 2008, from http://www.cs.ualberta.ca/ lindek/downloads.htm.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

M. Meila. 2003. Comparing clusterings. Techical Report TR418, University of Washington, Department of Statistics.

I. Mel'chuk and A. Zholkovsky. 1984. Tolkovo-kombinatornyj slovar'sovremennogo russkogo jazyka. *Opyty semantiko-sintaksiceskogo opisanija russkoj leksiki. Wien*.

I. Mel'chuk. 1974. Opyt teorii lingvisticheskix modelej tipa smysl¡==¿ tekst.

I. Mel'chuk. 1982. Lexical functions in lexicographic description. In *Proceedings, Eighth Annual Meeting of the Berkeley Linguistics Society*, pages 427–444.

I. Mel'chuk. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. *Lexical Functions in Lexicography and Natural Language Processing*, 31:37–102.

R. Mihalcea and P. Edmonds, editors. 2004. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, July. Association for Computational Linguistics.

R. Mihalcea, T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain, July. Association for Computational Linguistics.

G. Miller and W. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

R. Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia, July. Association for Computational Linguistics.

M.P. Oakes. 1998. *Statistics for corpus linguistics Edinburgh textbooks in empirical linguistics*. Edinburgh university press.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

M. Palmer, H. Dang, and C. Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

P. Pantel and D. Lin. 2000. An Unsupervised Approach to Prepositional Phrase Attachment using Contextually Similar Words. In *Annual Meeting – Association for Computational Linguistics*, volume 38(1), pages 101–108.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD02*.

P. Pantel. 2003. *Clustering by Committee*. PhD dissertation, Department of Computing Science, University of Alberta.

T. Pedersen. 1998. Dependent bigram identification. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence table of contents*. American Association for Artificial Intelligence Menlo Park, CA, USA.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of english words. In *Meeting of the Association for Computational Linguistics*, pages 183–190.

J Preiss and D. Yarowsky, editors. 2001. *Proceedings of the Second Int. Workshop on Evaluating WSD Systems (Senseval 2)*. ACL2002/EACL2001.

J. Pustejovsky and B. Boguraev. 1993. Lexical knowledge representation and natural language processing. *Artif. Intell.*, 63(1-2):193–223.

J. Pustejovsky and A. Rumshisky. 2008. Between chaos and structure: Interpreting lexical data through a theoretical lens. *Special Issue of International Journal of Lexicography in Memory of John Sinclair*, 21.

J. Pustejovsky, P. Hanks, and A. Rumshisky. 2004. Automated Induction of Sense in Context. In *COLING 2004, Geneva, Switzerland*, pages 924–931.

J. Pustejovsky, N. Calzolari, A. Rumshisky, J. Moszkowicz, E. Jezek, V. Quochi, and O. Batiukova. 2008a. Argument selection and coercion task. Semeval-2 task proposal, Brandeis University, Waltham, Mass.

J. Pustejovsky, A. Rumshisky, J. Moszkowicz, and O. Batiukova. 2008b. GLML: A generative lexicon markup language. Annotation guidelines, Brandeis University, Waltham, Mass.

J. Pustejovsky, A. Rumshisky, J. Moszkowicz, and O. Batiukova. 2009. GLML: Annotating argument selection and coercion. *IWCS-8: Eighth International Conference on Computational Semantics*.

J. Pustejovsky. 1995. *Generative Lexicon*. Cambridge (Mass.): MIT Press.

J. Pustejovsky. 2001. Type construction and the logic of concepts. In *The Syntax of Word Meaning*. Cambridge University Press, Cambridge.

J. Pustejovsky. 2005. A survey of dot objects. Technical report, Brandeis University.

J. Pustejovsky. 2006. Type theory and lexical decomposition. *Journal of Cognitive Science*, 6:39–76.

J. Pustejovsky. 2007. Type Theory and Lexical Decomposition. In P. Bouillon and C. Lee, editors, *Trends in Generative Lexicon Theory*. Kluwer Publishers (in press).

J. Pustejovsky. 2008. From concepts to meaning: The role of lexical knowledge. *Unity and Diversity of Languages*, pages 73–84.

P. Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

B. Roark and E. Charniak. 1998. Noun-phrase co-occurence statistics for semi-automatic semantic lexicon construction. In *COLING-ACL*, pages 1110–1116.

P. M. Roget, editor. 1962. *Roget's Thesaurus*. Thomas Cromwell, New York, 3 edition.

B. Rosario and M. Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

A. Rumshisky and J. Pustejovsky. 2006. Inducing sense-discriminating context patterns from sense-tagged corpora. In *LREC 2006, Genoa, Italy*.

A. Rumshisky, P. Hanks, C. Havasi, and J. Pustejovsky. 2006. Constructing a corpus-based ontology using model bias. In *The 19th International FLAIRS Conference, FLAIRS 2006*, Melbourne Beach, Florida, USA.

A. Rumshisky, V. A. Grinberg, and J. Pustejovsky. 2007. Detecting Selectional Behavior of Complex Types in Text. In P. Bouillon, L. Danlos, and K. Kanzaki, editors, *Fourth International Workshop on Generative Approaches to the Lexicon*, Paris, France.

A. Rumshisky. 2008. Resolving polysemy in verbs: Contextualized distributional approach to argument semantics. *Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica*.

J. Ruppenhofer, M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

J. Sinclair and P. Hanks. 1987. *The Collins Cobuild English Language Dictionary*. HarperCollins, 4th edition (2003) edition. Published as Collins Cobuild Advanced Learner's English Dictionary.

B. Snyder and M. Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

E. Velldal. 2005. A fuzzy clustering approach to word sense discrimination. In *Proceedings of the 7th International conference on Terminology and Knowledge Engineering*, Copenhagen, Denmark.

J. Véronis. 2004. Hyperlex: Lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

J. Weeds and D. Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

J. Weeds, D. Weir, and D. McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of CoLing 2004*, Geneva, Switzerland.

D. Widdows and B. Dorow. 2002. A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics*, pages 1093–1099, Taipei, Taiwan, August.

Y. Zhao and G. Karypis. 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55:311–331.

A. Zholkovsky, N.N. Leont'eva, and Y.S. Martem'yanov. 1961. O printsipial'nom ispol'zovanii smysla pri mashinnom perevode/on a principled use of sense in machine translation. *Mashinnyy perevod*, 2.

A.M. Zwicky and J.M. Sadock. 1975. Ambiguity tests and how to fail them. *Syntax and Semantics*, 4(1):1–36.