ABSTRACT

Title of Document: A COMPUTATIONAL THEORY OF THE USE-MENTION DISTINCTION IN NATURAL LANGUAGE

Shomir Wilson, Ph.D., 2011

Directed By: Professor Donald Perlis,
Department of Computer Science

To understand the language we use, we sometimes must turn language on itself, and we do this through an understanding of the use-mention distinction. In particular, we are able to recognize mentioned language: that is, tokens (e.g., words, phrases, sentences, letters, symbols, sounds) produced to draw attention to linguistic properties that they possess. Evidence suggests that humans frequently employ the use-mention distinction, and we would be severely handicapped without it; mentioned language frequently occurs for the introduction of new words, attribution of statements, explanation of meaning, and assignment of names. Moreover, just as we benefit from mutual recognition of the use-mention distinction, the potential exists for us to benefit from language technologies that recognize it as well. With a better understanding of the use-mention distinction, applications can be built to extract valuable information from mentioned language, leading to better language learning materials, precise dictionary building tools, and highly adaptive computer dialogue systems.

This dissertation presents the first computational study of how the use-mention distinction occurs in natural language, with a focus on occurrences of mentioned language. Three specific contributions are made. The first is a framework for identifying and analyzing instances of mentioned language, in an effort to reconcile elements of previous theoretical work for practical use. Definitions for mentioned language, metalanguage, and quotation have been formulated, and a procedural rubric has been constructed for labeling instances of mentioned language. The second is a sequence of three labeled corpora of mentioned language, containing delineated instances of the phenomenon. The corpora illustrate the variety of mentioned language, and they enable analysis of how the phenomenon relates to sentence structure. Using these corpora, inter-annotator agreement studies have quantified the concurrence of human readers in labeling the phenomenon. The third contribution is a method for identifying common forms of mentioned language in text, using patterns in metalanguage and sentence structure. Although the full breadth of the phenomenon is likely to elude computational tools for the foreseeable future, some specific, common rules for detecting and delineating mentioned language have been shown to perform well.

A COMPUTATIONAL THEORY OF THE USE-MENTION DISTINCTION IN NATURAL LANGUAGE

by

Shomir Wilson.

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:

Professor Donald Perlis, Chair
Professor Rance Cleaveland
Professor Norbert Hornstein
Professor Tim Oates
Professor Philip Resnik

For my parents

and

In loving memory of Warren Audus Wilson (1929-2011)
grandfather, patriarch

Special thanks to:
Scott Fults
Darsana Josyula
Michael L. Anderson
Matt Schmill
Josh Jones
Robert Dale
Min-Yen Kan
Elizabeth Traut
Jack Dudley

This work was funded in part by:
The Office of Naval Research (Grant #N000140910328)
The Air Force Office of Scientific Research (Grant #FA95500910144)
The National Science Foundation (Grants #IIS0803739,
#OISE1015666, #OISE0914091)
The Australian Academy of Sciences
The National Research Foundation of Singapore

Table of Contents

Publication Notes vi
Chapter 1: Overview 1
1.1 Introduction 1
1.2 Motivation 2
1.3 Dialog Systems 4
1.3.1 The Status Quo 4
1.3.2 The ALFRED Dialog System 9
1.3.3 Development of a New ALFRED Architecture 12
1.4 Hypothesis and Scope 13
1.5 Approach 15
1.6 Outline 16
Chapter 2: The Use-Mention Distinction 18
2.1 Introduction 18
2.2 Basic Concepts 18
2.2.1 Use-Mention Terminology 18
2.2.2 A Definition for Mentioned Language 21
2.3 Qualities of Mentioned Language 23
2.3.1 Preface to the List 23
2.3.2 List of Qualities 24
2.3.3 Practical Considerations 31
2.4 Relationship with Natural Language Processing 32
2.4.1 Part-of-Speech Tagging and Parsing 33
2.4.2 Conversational Systems 36
2.4.3 Lexical Semantics 37
2.4.4 Other Topics 38
2.5 Operationalizing the Definition 40
2.5.1 A Rubric for Mentioned Language 40
2.5.2 Examples of Rubric Usage 43
2.5.3 The Definition-Rubric Relationship 47
2.5.4 SRT Sentences 48
2.6 Conclusions 50
Chapter 3: Related Work 51
3.1 Introduction 51
3.2 Theories of the Use-Mention Distinction 52
3.2.1 Exclusive Theories 52
3.2.2 Non-Exclusive Theories 56
3.2.3 Applicative Theories 59
3.3 Natural Language Processing 61
3.4 Commonsense Reasoning About Language 63
3.5 Discussion 66
Chapter 4: Creation of a Robust Corpus of Mentioned Language 68
4.1 Introduction 68
4.2 The Pilot Study 68

4.2.1 Rationale for Choosing Wikipedia 69
4.2.2 Corpus Creation 73
4.2.3 Results and Discussion 75
4.2.4 Insights on Mentioned Language 76
4.2.5 Bridge to the Combined Cues Study 77
4.3 The Combined Cues Study 78
4.3.1 Candidate Collection and Labeling 79
4.3.2 Reliability and Consistency 81
4.3.3 Corpus Composition 82
4.3.4 Observations 86
4.4 The Enhanced Cues Study 87
4.4.1 Motivation 87
4.4.2 Changes from the Previous Studies 88
4.4.3 Corpus Creation 93
4.4.4 Annotation Results 95
4.4.5 Inter-Annotator Agreement Study 100
4.4.6 Discussion 102
4.4.7 Conclusions 104
Chapter 5: Detection and Delineation of Mentioned Language 105
5.1 Introduction 105
5.2 Motivation 105
5.3 Approach 108
5.3.1 Preliminaries 108
5.3.2 Detection Task 109
5.3.3 Delineation Task 110
5.4 Results and Discussion 113
5.4.1 Detection Task 113
5.4.2 Delineation Task 117
5.5 Conclusions 120
Chapter 6: Closing Thoughts 121
6.1 Discussion 121
6.2 Contributions 122
6.3 Future Work 123
Appendix A: Annotation Guidelines 125
Appendix B: Pseudocode of the Rubric 126
Appendix C: The Token-Type Distinction and the Definition 128
Bibliography 130

Publication Notes

Portions of Chapter 2 and Section 4.2 are from a paper published as

Distinguishing use and mention in natural language. Shomir Wilson. In *Proceedings of the NAACL HLT Student Research Workshop*, 29–33. Los Angeles, CA: Association for Computational Linguistics. 2010.

Portions of Chapter 2 and Section 4.3 are from a paper accepted by CICLING 2011 and published as

In search of the use-mention distinction and its impact on language processing tasks. Shomir Wilson. To appear in the *International Journal of Computational Linguistics and Applications*.

Chapter 1: Overview

1.1 Introduction

> *Every quotation contributes to the stability or enlargement of the language.*
>
> —Samuel Johnson (1709-1784)
>
> *I hate quotation. Tell me what you know.*
>
> —Ralph Waldo Emerson (1803-1882)

In order to understand language that we use, we sometimes must turn language on itself. British writer Samuel Johnson was not a scholar in natural language processing or computational linguistics, but, his above statement is surprisingly prescient. Through quotation, mention of language, and metalanguage—that is, language about language—we stabilize communication and keep it running smoothly in spite of the continuing evolution of language and the inevitable misunderstandings along the way. Emerson, in this unforeseen context, expresses a wry contradictory sentiment: one who learns through language must sometimes learn about language through its own explicit mechanisms. Often, quotation is how we say we know.

The use-mention distinction can be illustrated with a pair of sentences:

(1) The cat is on the mat.

(2) The word “cat” is spelled with three letters.

A reader easily understands that *cat* in the first sentence refers to an animal entity in a real or hypothetical world, while the same word in the second sentence refers to the word *cat*. The *use-mention distinction* is well-known and has a history of theoretical examination, but its actual patterns of appearance in natural language have received little study. Claims have been made on how humans so easily detect the distinction, but little (if any) previous work has been done to identify specific cues in language that enable this skill. Instances of mentioned language are easy to conjure, but never previously have they been collected in large numbers for aggregate study.

This dissertation will begin to address these gaps in our understanding, for the benefit of computer applications that must process and learn from natural language. This chapter will provide a brief overview of the motivation, goals, and approach of the project.

1.2 Motivation

The historical lack of attention to the use-mention distinction might suggest that it is peripheral to the study of language, but this is far from the truth. Evidence suggests that human communication frequently employs the use-mention distinction, and we would be severely handicapped without it (Perlis, Purang, and Andersen 1998). In both written and spoken contexts, the mention of letters, sounds, words, phrases, or entire sentences is essential for many language activities, including the introduction of new words, attribution of statements, explanation of meaning, and assignment of names (Saka 1998). Moreover, detecting the distinction is a nontrivial task. While stylistic cues like italic text or quotation marks are sometimes used to indicate the mention of language, such cues are not applied (or available) equally in all contexts. Even when they are applied uniformly, they tend to be “overloaded” with other uses as well (e.g., emphasis). Cues such as pauses and gestures exist in spoken conversation, but these too are only approximate indicators and are easily lost in transcription.

Just as humans benefit from mutual recognition of the use-mention distinction, the potential exists for us to benefit from language technologies that can recognize it as well. With a model of the mechanics of mentioning language, the following will become possible:

- Dialog systems can be designed to recognize when a user is attempting to correct misinterpreted statements or introduce new terms, instead of ignoring or misinterpreting those activities. This was a prime motivation of this work, and contributions toward creating such a dialog system will be discussed in detail in the next section.
- Lexical semantics tools can take advantage of (or assign greater weight to) information encoded in mentioned language, since it tends to be direct, salient, and unambiguous.
- Trends in language can be studied with special attention to how new terms are effectively (or ineffectively) introduced.
- Source attribution tools can be trained to precisely identify which words are being reproduced in a quotation without the aid of stylistic cues, such as quotation marks or italics.

- Language learning materials (especially for second language acquisition) can be prepared with special attention to strategies in metalanguage that have been found to be most effective.
- Typesetting and copyediting software can be designed to recognize instances of mentioned language and apply stylistic features to them uniformly.

Thus, advancing this area of knowledge could benefit several lines of research in computational linguistics, natural language processing, and artificial intelligence.

1.3 Dialog Systems

This section presents some motivation for the dissertation from research in dialog systems. A contribution in the form of continuing work on ALFRED, a dialog system that explicitly represents and reasons about language knowledge, is also discussed.

1.3.1 The Status Quo

A *dialog system* is a computer system that converses with a human via natural language. Dialog systems can be applied to process communication between a user and a domain (e.g., an information source or a controllable device) when conversation (either written or spoken) is a desirable mode of interaction (Josyula 2005; Lester, Branting, and Mott 2004). A sufficiently flexible dialog system can ease the user's "learning curve" when interacting with a new system (Litman and Pan 2002) or enable human-computer interaction in situations where speech is the only available channel of communication, such as over the phone (Raux et al. 2005). Figure 1.1

shows the role of a dialog system as an intermediary between a human user and a domain.

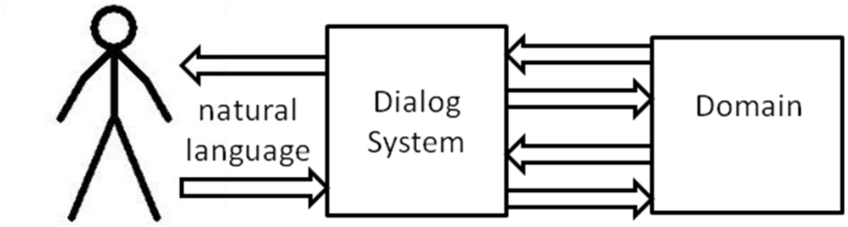

Figure 1.1: The basic model of interaction between a human, a dialog system, and a domain.

The human and the dialog system communicate bidirectionally, taking turns with utterances in natural language. The dialog system also communicates (either bidirectionally or unidirectionally) with the domain using a suitable formal protocol. Through this model, the user can affect changes in the domain or access information in it using natural language. Dialog systems are frequently *task-oriented* (Josyula, Anderson, and Perlis 2003): they are designed to cooperate with users to perform specific activities or achieve certain outcomes.

Task-oriented dialog requires a dialog system to have some knowledge of the interactive components of the domain. However, the explicit representation of *language* knowledge has generally received little attention in research on dialog systems (Anderson et al. 2002). This has led to many systems with knowledge bases which resemble the fragment shown in Figure 1.2 below. The emphasis in this typical model is on domain knowledge, with relatively sparse representation of language knowledge. Language knowledge is rigidly structured, static to the user, and linked only minimally to domain knowledge. The language knowledge is not designed to address all concepts in the domain or to stand on its own as a potential domain to be reasoned about, and thus its representation is sparser.

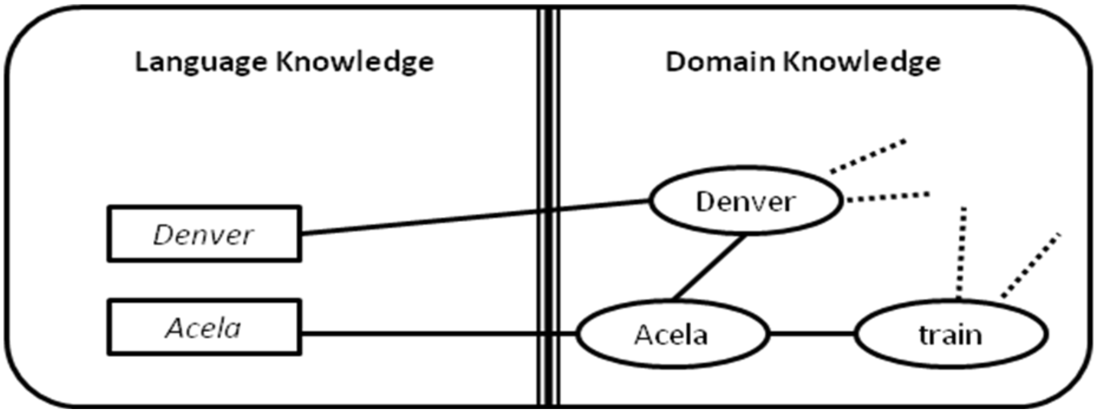

Figure 1.2: Example fragment of a knowledge base of a typical dialog system, using trains and cities as an example domain. [1]

Having knowledge about language and being able to reason about it are core components of *conversational adequacy* (Perlis, Purang, and Andersen 1998), the ability to engage in flexible, robust conversation. Humans possess conversational adequacy and frequently make use of it in dialog. Figure 1.3 below presents some exchanges in dialog that illustrate this quality.

These illustrate several different functions of mentioned language:

- In Exchange (1), Partner A asks Partner B to clarify the spelling of his name, which is not present in his utterance but is presumably present in the language knowledge of both participants. To do this, he mentions a letter of the alphabet, as does Partner B, correcting him.

[1] Traum et al. (1996) originally introduced this domain in several studies of human-human and human-computer dialog.

- In Exchange (2), Partner A asks partner B for a word to refer to the object specified by the deictic “that”; when Partner B responds, she introduces the term “dudwidler”.
- In Exchange (3), Partner A refers to the conversation history to clarify a missed or misunderstood utterance. Partner B mentions her previous utterance.
- In Exchange (4), Partner A asks Partner B to clarify pronunciation of his name. Partner B pronounces the name with emphasis, then mentions one of the syllables, and finally provides the pronunciation of another word as an example, mentioning it as well.
- In Exchange (5), Partner A asks for the grammatical category of a word, creating a disfluent sentence in which “is” is used and subsequently mentioned. Partner B answers with the same disfluency.

1) A: Do you spell your name with one or two “z”’s? B: With one ‘s’.
2) A: What’s that called? B: It’s called a “dudwidler”.
3) A: Wait, what did you say? B: I said, “I need the orange parts tomorrow”.
4) A: How do you pronounce your name? B: You say it “JAY-son”. “JAY” as in “blue jay”.
5) A: Is “is” a verb”? B: Yes, “is” is a verb.

Figure 1.3: Five fragments from (separate) dialogs that show conversational adequacy, between human dialog partners labeled “A” and “B”.

Because of its many uses and its essentiality to communication, conversational adequacy is easily taken for granted when speaking with a person. However, the *lack* of conversational adequacy demonstrated by computer dialog systems can make normal conversation difficult or impossible. Figure 1.4 illustrates this with fragments from conversations with two such systems. The first, named *Let's Go!* (Eskenazi et al. 2008), is a system designed to help users plan trips on the Pittsburgh bus system. Users speak with it via telephone, and speech recognition errors occur when audio quality is poor (or even sometimes when it is not). The second dialog system, called *Suzette*, was the 2010 winner of the Loebner Prize, an annual contest that gives awards to the most human-like dialog agents[2]. This accolade might lead one to believe that Suzette is a standout among her computerized peers in conversational adequacy, but the fragment shows otherwise.

```
Let's Go!: Where do you wish to depart from?
User: Arlington.
Let's Go!: Departing from Allegheny West. Is this right?
User: No, I meant "Arlington".
Let's Go!: Please say where you are leaving from.
```

```
User: Do your friends call you "Suzy" or "Suzette"?
Suzette: I'm crazy.
User: Did you understand my question?
Suzette: How much do you weigh?
```

Figure 1.4: Fragments of human-computer conversations with *Let's Go!* and Suzette, two dialog systems.

In the *Let's Go!* fragment, the user says that he wishes to depart from Arlington, but the dialog system mishears it as "Allegheny West". The system asks for confirmation,

[2] Detailed results of the 2010 competition are available at `http://loebner.net/Prizef/2010_Contest/results.html` .

and the human tries to correct the misunderstanding by mentioning his previous utterance. However, *Let's Go!* does not recognize the correction and reverts to a rewording of its original question. In the *Suzette* fragment, the user asks the system whether its friends refer to it as "Suzy" or "Suzette" by mentioning the two names. *Suzette* dodges the question entirely, and remains evasive when the user checks if his previous utterance was understood. In both of these conversations, a human dialog partner in place of the computer system would easily understand the user. However, *Let's Go!* cannot process a simple question response containing mentioned language, and *Suzette* shows complete incapability to discuss mentioned language in the form of names or conversation history.

The behaviors of *Let's Go!* and *Suzette* when faced with mentioned language or metalanguage are not unusual among dialog systems. The status quo in dialog systems research is occupied almost entirely by systems that share the same deficiencies. These include (but are not limited to) RavenClaw (Bohus and Rudnicky 2009), SNePS (SC Shapiro and Kandefer 2005), Basilica (Kumar and Rosé 2009), and TRIPS (Blaylock et al. 2010). However, one effort to address this problem is the ALFRED project, described in the next subsection.

1.3.2 The ALFRED Dialog System

ALFRED (an acronym: Active Logic for Reason-Enhanced Dialog) is a task-oriented dialog system built to explore how such a system can use metalanguage and metareasoning to exhibit conversationally adequate behavior (Anderson et al. 2008). ALFRED serves as a universal interfacing agent for the user to control a variety of simulated domains, such as trains on a track system, lights in a house, a pool thermostat system, and a media player. The user engages in mixed-initiative dialog with ALFRED to accomplish tasks. The user may take the initiative by issuing commands to domains or asking questions about their status. ALFRED may take the initiative through metalinguistic dialog, by requesting that the user clarify the content of a command or question, thus stabilizing the conversation when misunderstandings occur or new words appear. To decide when to initiate reparative dialog, ALFRED uses a form of the metacognitive loop (Anderson et al. 2007) to *note* when an anomaly occur in conversation, *assess* the cause of a problem, and *guide* a solution into place.

Figures 1.5-6 below show the evolution of a fragment of ALFRED's knowledge base during a dialog with metalinguistic content[3]. The fragment is modeled after the TRAINS domain from the previous subsection.

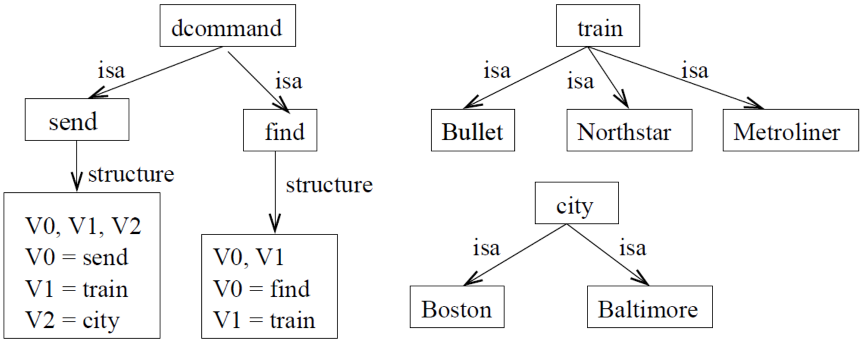

Figure 1.5: A fragment of ALFRED's knowledge base prior to the user utterance "Send the Subway to Boston".

[3] This example appears in greater detail in a paper (Josyula et al. 2007) which the present researcher co-authored.

The user begins with the command “Send the Subway to Boston”. The *send* command is one that ALFRED recognizes and can parse; however, the train *Metro* is not in his knowledge base, which contains only *Bullet*, *Northstar*, and *Metroliner*. Instead of rejecting the command altogether, ALFRED *notes* that it is unable to understand the full utterance and *assess* that the single word “Subway” is problematic, and moreover, that “Subway” should identify a train. ALFRED then *guides* a solution into place by taking the initiative to ask the user “Which train is Subway?” If the user answers “Subway is Metroliner” (or even simply “Metroliner”), ALFRED amends its knowledge base, as shown in Figure 1.6 below.

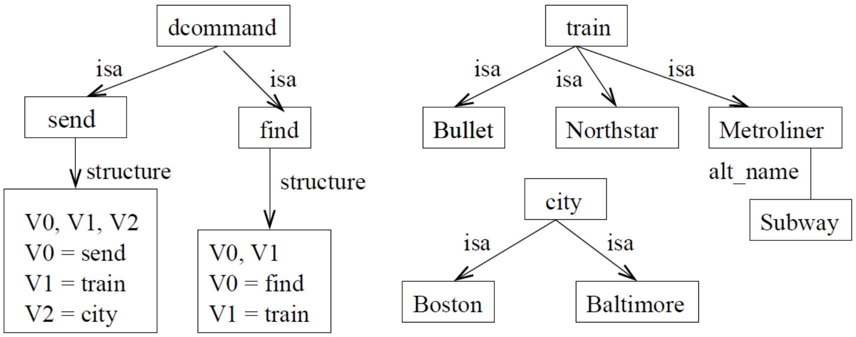

Figure 1.6: A fragment of ALFRED’s knowledge base following the user utterance “Subway is Metroliner”.

Subway is now represented in ALFRED’s knowledge base as an alternate name for *Metroliner*. The metacognitive loop is complete, and the system returns to interpreting the user’s original utterance. ALFRED sends Metroliner to Boston and informs the user of the result of this action.

1.3.3 Development of a New ALFRED Architecture

An important contribution of the ALFRED system as described in the above subsection is its ability to engage in metareasoning about language. However, another step towards conversational adequacy is the ability to engage in explicit language-mention, and a new knowledge architecture was implemented for this purpose. Figure 1.7 below shows a fragment of ALFRED's new *concept space*, again focusing on TRAINS.

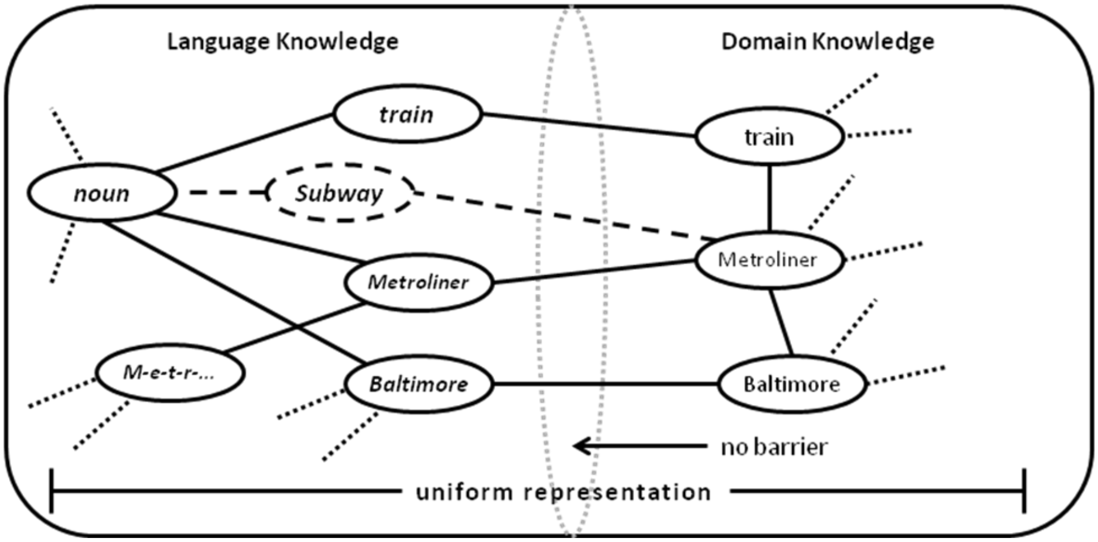

Figure 1.7: A fragment of ALFRED's new knowledge base, shown in a manner to contrast with Figure 1.2.

The major difference between this architecture and that of other dialog systems (shown in Figure 1.2) is the explicit, structured representation of language knowledge. Domain knowledge and language knowledge are now represented uniformly, allowing ALFRED to apply the same reasoning facilities to both. This uniformity also allows language knowledge to be flexibly structured and open to change during user interaction. For example, Figure 1.7 illustrates the word

“Subway” being added to the concept space as an alternate name for Metroliner, part of the narrative in the previous subsection. The word “Subway” is not only linked to the train *Metroliner* but also to the word category *noun*, permitting ALFRED to reason about nouns and their role in language. The user is also now able to speak about an indeterminate train or all trains, since a representation of the word is present in the system’s language knowledge. Moreover, spelling is now represented in the concept space, further facilitating language-mention through speech recognition front-ends attached to ALFRED.

The explicit representation of language knowledge in a dialog system is a step toward conversational adequacy, but another important step remains missing. To process mentioned language “naturally”, ALFRED must recognize the linguistic cues that signal when users are mentioning language and determine where specifically in an utterance the mention occurs. The need to fill this gap was one of the core motivations of this dissertation.

1.4 Hypothesis and Scope

This dissertation will examine the use-mention distinction with a focus on detecting and delimiting *mentioned language*. Initial, informal observations suggest that mentioned language tends to occur either less often or less prominently than used language, and it tends to be a phenomenon identified by and surrounded by used language. Additionally, although the breadth and complexity of mentioned language will be discussed fully in Chapter 2, some restrictions in scope will be necessary to make the computational aspects of the dissertation feasible. Two of these are discussed below.

First, *written language* will be the preferred medium for studying the use-mention distinction. Although some of the applications of use-mention detection involve conversational language, it was decided to focus on written language first, for its relative consistency and ease of analysis. If metalanguage competency is truly a core language skill (Anderson et al. 2002), it is likely that some or many features of mentioned language will remain static across different communication media, and future research efforts may test this.

Second, explicit instances of mentioned language will be the focus of detection efforts. The mention of language many occur implicitly in variety of language phenomena, including irony and emphasis (Saka 2003; Sperber and D Wilson 1981). Some existing theories of the distinction, discussed in Chapter 3, allow for the frequent—or even ubiquitous—coexistence of use and mention as aspects of communication. Proper evaluation of such theories will be beyond the scope of this dissertation (although the researcher believes them to be more than plausible), which will focus instead on explicit mention of language, due to its tractability as a computational problem in the present state of language technologies.

Within these parameters, this dissertation will examine the hypothesis that cues in vocabulary, sentence structure, and semantic roles will be sufficient to identify most instances of mentioned language. Identification of an instance will consist of two activities:

- *Detection: Determining whether mentioned language is present in a given string of words*: Such strings of words will be individual sentences. For the parameters described above, this is a binary decision: either a sentence contains mentioned language, or it does not.

- *Delineation: Determining which words are mentioned*: In each sentence that contains mentioned language, a sequence of words must be identified that are being mentioned. In some instances, these words might simultaneously be used language as well.

1.5 Approach

Three tasks, listed below, will be necessary to study the use-mention distinction and test the hypothesis.

(1) A conceptual framework must be established for examining mentioned language. Previous theoretical treatments of quotation and the use-mention distinction often disagreed on the terminology, qualities, and pervasiveness of the phenomenon, and thus supply very few "ground rules" for an empirical study of it. The conceptual framework presented in this dissertation will be as inclusive of prior work as possible while being consistent with the practical nature of the hypothesis.

(2) Mentioned language must be studied empirically. Since little prior work exists on the topic, this task will involve creating corpora of text with labeled instances of mentioned language, so that their properties can be examined in aggregate. A combination of automated and manual techniques will be used to gather instances, since neither alone will be sufficient: human readers are impractically slow, but at this stage it will not be possible to detect the phenomenon without human intervention.

(3) Techniques will be developed to identify instances of mentioned language computationally, without human intervention. Some strategies for doing this will first appear in corpus analysis, and once the corpora are complete it will be possible to directly address the hypothesis.

Beyond this dissertation, the corpora and detection techniques will be available for future research on both the use-mention distinction and related topics.

1.6 Outline

These three tasks—building a conceptual framework, constructing corpora, and enabling computational identification—will be addressed in the four core chapters of this dissertation. Figure 1.1 below illustrates how the dissertation structure and the tasks relate.

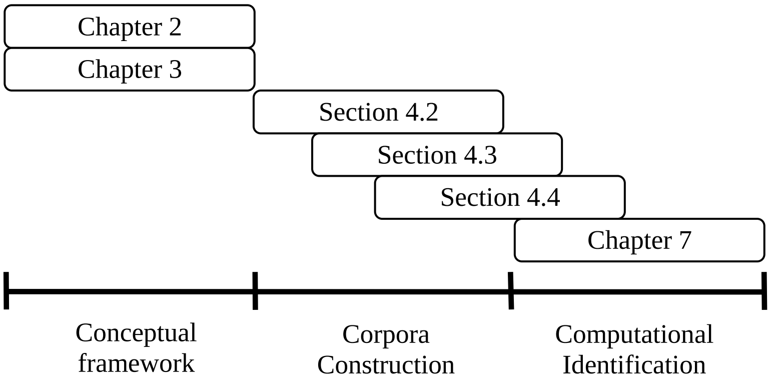

Figure 1.1: The relationship between the dissertation structure and tasks.

Chapter 2 will describe the use-mention distinction in detail, introducing necessary terminology and definitions. An operational characterization of the distinction will be introduced, along with many illustrative examples. A list of categories of mentioned language will be constructed. The effects of mentioned language on language processing tasks will be explained. Chapter 3 will review previous work on the use-mention distinction and some related topics.

Chapter 4 will describe three corpora of mentioned language, constructed using progressively more sophisticated methods. *Wikipedia* will be identified as a uniquely suitable source of text, and *stylistic cues* and *mention words* will be introduced as guides for sifting through text to find candidate instances. Sections 4.3 and 4.4 will begin to identify practical methods for automatic identification of mentioned language in any text.

Chapter 5 will address the problem of computational identification. A combination of machine learning and hand-coded rules will be explored, and some performance limitations of mentioned language detection will be discussed. Finally, Chapter 6 will outline some final thoughts and future work.

Chapter 2: The Use-Mention Distinction

2.1 Introduction

This chapter will accomplish the following:

- The use-mention distinction will be introduced, along with definitions for the related terms *mentioned language, metalanguage,* and *quotation*;
- Qualities of mentioned language will be listed, to demonstrate the breadth and complexity of the phenomenon;
- Those qualities will be reviewed for their practical applicability to this study;
- Beneficial relationships will be predicted between detecting mentioned language and several topics in natural language processing research; and
- A rubric for detecting mentioned language will be proposed, to provide consistency and objectivity when hand-labeling instances of the phenomenon.

2.2 Basic Concepts

2.2.1 Use-Mention Terminology

Although the reader is likely to be familiar with the use-mention distinction, the topic merits further explanation to establish what precisely is being referred to. Intuitively, the vast majority of language is produced for use rather than mention, as the roles of mention are auxiliary (albeit indispensible) to language use. For brevity, this dissertation will adopt the terms *language-mention* to refer to the act of mentioning language and *mentioned language* to refer to linguistic entities produced for the purpose of mentioning them. The terms *language-use* and *used language* will appear occasionally and will carry the expected complementary meanings.

The use-mention distinction, as one might expect, is the distinction between using linguistic entities (such as letters, symbols, sounds, words, phrases, or sentences) and mentioning them. Since this explanation is slightly opaque at best, some examples and a proposal for a definition will follow. Consider example sentences from Section 1.1, reproduced below:

(1) The cat is on the mat.

(2) The word "cat" is spelled with three letters.

In (1), the reader's attention to meaning does not focus on the words themselves, but instead on the presumed cat on the mat. We say the word "cat" in particular is being *used* (to refer to something other than itself, namely to a kind of animal, not even a word at all) in (1); and so is the word "mat". This can be taken, perhaps, as the standard way any word is employed in sentences: to call attention to something beyond the mere word. In (2), the reader understands that it is the word *cat*—a string of three letters, as opposed to any particular cat or cat concept—that is in the focus of the sentence. In such a case we say the word is being mentioned, not used. Quotation marks around *cat* in (2) are a convention to further reinforce that the word is being mentioned, and in some contexts (such as this sentence) italics may serve the same purpose.[4] "Setting aside" mentioned language via stylistic cues is a common

[4] In a manner of speaking, quotation marks provide a *name* that refers to the quoted word. Hence the quotes-plus-word as a unit is being *used* to refer to the word inside quotes, and the word itself remains a case of mention. This is not a serious problem, but it foreshadows an issue that will arise later. This chapter will follow convention, however, and speak of a quoted-expression unit as a case of mentioned language (of the item inside quotation marks).

convention, as reflected in popular style guides (Jr. Strunk and White 1979; Chicago Editorial Staff 2010) for formal writing. In spoken language, nonverbal cues are often present to delimit mentioned language, such as prosodic features (e.g., intonation, stress) or gestures.

The other linguistic entities listed above can also be mentioned; for example:

(3) The Classical Latin alphabet did not contain a ‘w’.

(4) Mathematical symbols, such as ‘∞’, are available for some fonts.

(5) The rusty hinge emitted a sharp “eeeeek” sound as it closed.

(6) “Behind the eight” is an idiom that originated from the game of billiards.

(7) The sentence “The cat is on the mat” appears in many linguistics papers.

Longer linguistic entities (such as paragraphs) are also subject to language-mention, though this occurs less frequently and places a greater burden on an audience’s understanding of the phenomenon. One frequent sentence-length role of language-mention is *quotation*, in which language from another source is reproduced as part of a statement, as in (8) below:

(8) Eric said, “We should meet for lunch.”

In (8), the phrase between quote marks is mentioned as what Eric has said. However, a reader is likely to react to the quoted text as a string with semantic depth, indicating that the *use* disjunct of the use-mention distinction is present as well. This mix of use and mention is common in quotation, as we tend to quote linguistic entities that possess meaning.

By necessity, discussions of language always invoke *metalanguage*, which is language used when language itself is being described (Audi 1995). Within the context of formal languages, metalanguage tends to be distinct and separate from *object language*—that is, the language being discussed. However, in natural language this separation does not hold; we use natural language to talk about natural language. Notably, sometimes two different natural languages are present when language-mention occurs, as in

(9) The French word *chat* refers to a feline animal.

Although English resembles a metalanguage in (9), it is clearly not a metalanguage in the general sense, as English is used for many "non-meta" functions (and French speakers can discuss English too). To retain some intuitions on the term *metalanguage* and also satisfy these practical constraints, within this dissertation it will refer chiefly to the words and syntactic structures that "frame" mentioned language in a sentence or provide linguistic cues for its presence.

2.2.2 A Definition for Mentioned Language

In spite of the ubiquity of the phrase *use-mention distinction*, it is difficult to find an explicit definition for either the distinction itself or its two disjuncts. The effort here will be to define *mentioned language* since it is less common, more peculiar, and far less studied than used language. Gestures toward definitions in previous literature, although vague, will be considered in the next chapter. For present purposes, the definition below will specifically cover *sentential* mentioned language, where the mentioned linguistic entity is referred to inside of the same sentence that it occurs. An example of a sentence that fails this requirement is:

(10) Disregard the last thing I said.

This restriction is necessary to reduce the complexity of the computational tasks that will follow in later chapters, and it will be assumed in further discussions unless explicitly stated otherwise. Also, although this definition is nominally applicable as a test to determine whether a token qualifies as mentioned language, it is not necessarily intended for that activity. An alternative mechanism for labeling candidate instances will follow in Subsection 2.5 in the form of a rubric, which will be easier for annotators to use when creating the corpora discussed in later chapters. A brief attempt to train annotators to use the definition was unsuccessful; hence the rubric was necessary in order to have an applicable mechanism for mention-detection.

Definition: For T a token or a set of tokens in a sentence, if T is produced to draw attention to a property of the token T or the type of T, then T is an instance of mentioned language.

Here, a *token* is the specific, situated (i.e., as appearing in the sentence) instantiation of one of the linguistic entities listed in 2.2.1—letters, symbols, sounds, words, phrases, or entire sentences (subsumed in longer sentences as independent clauses, as in (8) above). A *property* might be a token's spelling, pronunciation, meaning (for a variety of interpretations of that term), structure, connotation, original source (in cases of quotation), or another aspect for which language is shown or demonstrated. The *type* of T is relevant in some instances of mentioned language (such as in (2)) and the *token* itself is relevant in others, as in

(11) “The” appears between quote marks in this sentence.

Constructions like (11) are unusual and are of limited value in practical language use; the definition accommodates them for completeness. An extended discussion of the role of the token-type distinction in mentioned language can be read in Appendix C.

2.3 Qualities of Mentioned Language

2.3.1 Preface to the List

The following subsection contains a list of qualities of mentioned language that illustrate its structures, meanings, and roles in language. This list is a compilation of qualities from previous literature (when cited) and qualities that are readily evident in the given examples. Some qualities will be widespread or universal, while others will be intermittent or unusual. A few caveats apply to the list:

- The list is intended primarily for the properties of English language-mention. Although some items in the list might be applicable to other languages, further study will be necessary to determine their cross-language extent and whether any are truly universals.
- Although efforts were made toward coverage and completeness, the list is not purported to be a comprehensive inventory of qualities. Instead, it should be treated as a series of illustrations of the complexity and broad scope of the phenomenon.

- Not all of the qualities listed will be detectible by the computational methods discussed later in this dissertation. Some are difficult to detect consistently within the constraints of the hypothesis (see 1.3), and some are easily detectible by humans but require language skills that we have not yet been able to give to computers. Still others are unusual cases that, while perhaps detectible, have limited practical value.
- Disagreement exists over whether some of the listed qualities are valid for mentioned language. In the literature on the use-mention distinction, it is rare to find *explicit* exclusion of any of them; still, no single theory has accounted for all qualities. The list will be as inclusive as possible.

2.3.2 List of Qualities

In the example sentences that follow, either pairs of asterisks or quotation marks will be used to delimit instances of mentioned language. The qualities are organized into three categories: surface variations, quotational variations, and reference and semantics.

Category #1: Surface Variations

A. Syntactic Variety: Mentioned language generally takes the role of a noun phrase, though with exceptions shown later in this list. As a noun phrase, it can fulfill several different syntactic roles in a sentence. Three examples of this are:

(12) Two fictional superheroes are named *Gambler*.

(13) *Gambler* is the name of two fictional superheroes.

(14) The name *Gambler* is shared by two fictional superheroes.

In (12) *Gambler* appears as the object of a verb phrase; in (13) it appears the subject of a verb phrase; in (14) it appears in apposition with *name*.

B. Variety in Vocabulary: The vocabulary of metalanguage is flexible, and often a language user can choose among multiple words to frame an instance of mentioned-language:

(15) The child was named *Peter* after his father.

(16) The child was called *Peter* after his father.

(17) The child was christened *Peter* after his father.

While *named*, *called*, and *christened* have subtly different connotations, all three sentences succeed in assigning the designation *Peter* to the child in the context.

C. Explicitness: Framing metalanguage is common but varies in explicitness. For example, the two sentences below can share the same meaning:

(18) The word *go* appears on the screen after five minutes.

(19) *Go* appears on the screen after five minutes.

On the other hand, some cases of mentioned language require either appropriate world-knowledge or stylistic cues to detect. For instance:

(20) The teacher wrote “in the greenhouse” on the chalkboard.

suggests that the teacher wrote the exact three words *in the greenhouse* only by virtue of the quotation marks in around it. Without them, the audience would likely assume that the teacher was writing on a chalkboard positioned in a greenhouse. Similarly, when speaking of the popular children's television show *Sesame Street*,

(21) *Elmo* has four letters.

could mean that the name *Elmo* has four letters or the energetic red creature Elmo is holding four letter-shaped objects. (The former interpretation would have seemed likely if quotation marks had been used around *Elmo*.)

D. Stylistic and Paralinguistic Cues: As mentioned in 2.2.1, stylistic cues are sometimes used to delimit mentioned language. Three common cues in written language are bold text, italic text, and text between quotation marks:

(22) This is why the club is nicknamed ***The Jurists***.

(23) This is why the club is nicknamed **The Jurists**.

(24) This is why the club is nicknamed "The Jurists".

The particular choice of stylistic cue depends on convention, level of formality, and media (i.e., bold or italics are not always options). These three cues are unavailable in spoken language, and instead delimiters such as intonation, stress, or hand gestures are sometimes used.

E. Disfluency: Sometimes a mentioned linguistic entity does not usually appear as a noun, but language-mention gives it the qualities of one:

(25) "Has" is a conjugation of "have".

(26) The only word on the paper was "before".

These examples might cause a human reader to pause and reconsider while reading, but ultimately their meanings are clear.

Category #2: Quotational Variations

The following four qualities were identified by Cappelen and Lepore (1997).

F. Pure Mention: This is the “classic” form of mentioned language that all theories agree upon. No quotation is involved, and instead a statement is made about a property of a linguistic entity. Many examples of this have already been presented for short, noun phrase-like linguistic entities, but longer entities also can be mentioned purely:

(27) “The cat is on the mat” is a sentence.

This sentence does not assert that “The cat is on the mat” has been uttered before by any language user (although it has been many times). It is merely a statement about its acceptability as a sentence.

G. Direct Quotation: Utterances can be reported with framing metalanguage and stylistic cues which suggest precise reproduction:

(28) Baljeet said “The cat is on the mat”.

For (28) to be true, Baljeet must have said “The cat is on the mat”, with those exact words. Linguistic entities that have little or no semantic value can also be directly quoted:

(29) “_U2E+ha4” was scrawled upon the wall.

The string *_U2E+ha4* in (29) is also notable as an instance of quotation that has limited semantic depth, as it appears to be a random string of characters.

H. Partial Quotation[5]: Utterances can be reported with framing metalanguage and stylistic cues which suggest that only part of a statement is faithfully reproduced:

(30) Baljeet said the cat is “on the mat”.

For (30) to be true, Baljeet must have said that the cat is “on the mat”—though it is only necessary for those words between quotation marks to be his. He might have said, for instance: “Ted’s Abyssinian kitten is on the mat.” In both (28) and (30), the particular placement of the quotation marks supplies information to the audience, as they indicate exactly which words are faithfully reproduced.

I. Paraphrase: Although this quality does not satisfy the definition of mentioned language presented in 2.2.2, an utterance can be mentioned without *any* direct reproduction:

(31) Baljeet said the cat is on the mat.

In this case, the *absence* of stylistic cues is significant: for (31) to be true, Baljeet must have said that the cat is on the mat, though with an undetermined choice of words. The sentence mentions his utterance only indirectly, and it is not reproduced.

[5] Cappelen and Lepore use the term *mixed quotation* for this quality, but Maier uses *mixed quotation* to refer to the mixture of use and mention discussed in 2.2.1. *Partial quotation* is used here to avoid overloading the term.

Category #3: Reference and Semantics

J. Applicability to Medium: Some forms of mentioned language are of limited value when reproduced outside of their original media. For instance, the sentence

(32) *Kampung* is spelled *K A M P U N G*.

might be useful when spoken out loud but presents redundant information when written down, as it appears above. Conversely,

(33) That jingle *Dum, dum dum dum* sounds familiar.

conveys a rhythm when written down but does not convey the variations in pitch, forming a melody, which can occur when spoken out loud.

K. Explicit Mixed Referent: Sometimes a word or phrase is simultaneously used and mentioned. Consider the first word to appear between asterisks in

(34) *Color*, also spelled *colour*, is a visual perceptual property.

The phrase *also spelled colour* refers to *color* chiefly as a word (mentioned language), while *is a visual perceptual property* does not (used language).

L. Mixed Referent through Implicature: A mix of use-reference and mention-reference can also happen without any explicit cues in language. Consider a spoken dialogue between people standing in a circle: a participant might introduce someone new to the circle to those already participating by uttering

(35) Here’s John.

If all participants know who John is and all participants know his name, then (35) is a case of mere use, and it could be restated (albeit awkwardly) as

(36) Here's John the person.

However, if John is unknown to one or more participants, the speaker of (35) could have intended to introduce both the person John and the association between the name *John* and the new arrival. Then (35) could be restated as

(37) Here's the person named John.

Speakers understand "mixed" use-mention reference through *implicature* (Grice 1975), as it is neither explicitly stated nor logically implied by the sentence.

M. Ostention: Tokens of language can be mentioned with the intent of illustrating one of several different properties that they possess, such as orthographic form, lexical entry, phonic form, intension, and extension. These aspects have been termed *ostentions* (Saka 1998), and they will be discussed in greater detail in Chapter 3.

N. Self-Reference: While uncommon in practical use, it is possible for a sentence to mention itself, as in the two examples below.

(38) *This sentence has five words.*

(39) *This sentence is an example in a dissertation.*

This sentence in both (38) and (39) refers to the sentence in which it occurs. While (38) refers only to a surface feature of itself, (39) requires an evaluation of its context to determine its truth. Both tend to cause a reader to pause, and they present challenges to formal representation. Sentence (11) above is also an example of this.

O. Irony and Distancing: Ironical statements sometimes draw attention to the meaning of a specific word or phrase, in order to highlight its discord it with accepted circumstances. Consider, for instance, this utterance if said by a person walking through pouring rain:

(40) What *lovely* weather we are having.

Although *lovely* has not been *explicitly* mentioned by (40), some accounts of irony (Jorgensen, GA Miller, and Sperber 1984; Sperber and D Wilson 1981) hypothesize that the use-mention distinction is responsible.

2.3.3 Practical Considerations

The list in the preceding section draws from a variety of topics in syntax, semantics, and pragmatics. This dissertation focuses on the *detection* of mentioned language, since it is fundamental to further computational studies of the phenomenon, and some qualities in the list will receive greater attention than others. Listed below are the relationships between the qualities and the present study.

- Attempts will be made to account for syntactic variety (A) and variety in vocabulary (B) in the detection of mentioned language.
- Detection will focus on explicit instances (C), since implicit mentioned language often requires substantial word knowledge to detect.

- Stylistic cues (D) will be exploited to make the collection of instances of mentioned language a practical problem, but the study will strive to detect the phenomenon without such cues. Paralinguistic cues (D) will not be studied.
- Instances of the phenomenon that contain disfluency (E) will be detected whenever possible, though they may pose problems to detection methods.
- Pure mention (F) and direct quotation (G) will be included in this study.
- Partial quotation (H), however, is too difficult (if not impossible) to properly delimit without stylistic cues, and will not be addressed. Paraphrase (I) will be disregarded because it is unclear how to delimit it in a consistent, reproducible manner.
- Applicability to the medium (J) is not expected to be a substantial problem in detecting mentioned language, though its combination with other qualities might make it so.
- Explicit mixed referent (K) will be addressed, but the complexity in modeling implicature will make the automatic detection of (L) unfeasible.
- Ostentions (M) will receive further examination, since they provide a framework for examining the information that mentioned language conveys.
- Self-reference (N) and irony (O) will not receive further attention.

2.4 Relationship with Natural Language Processing

Thus far, the use-mention distinction has received relatively little consideration from natural language processing and artificial intelligence research in general. This section will explain the importance of detecting mentioned language and how some active areas of research stand to benefit. Section 2.4.1 will discuss the difficulty that mentioned language poses to part-of-speech taggers and parsers. However, parsing is not an end goal in itself, and Sections 2.4.2-4 will discuss how the detection of mentioned language can impact other automated language tasks.

2.4.1 Part-of-Speech Tagging and Parsing

Current part-of-speech taggers and parsers are agnostic to the difference between use and mention. One problem this has created is a lack of conventions on how to tag and parse mentioned language. An instance of mentioned language ostensibly functions as a noun, and most instances are labeled as such, simply because they tend to be nouns when they appear in language use. However, sequences of words that bear little resemblance to noun phrases (such as independent clauses, which appear when quoting speech acts) are equally instances of mentioned language. Existing corpora, to the knowledge of the writer, do not label mentioned language, though some make concessions toward them. The Penn Treebank, for instance, has tags for foreign words (FW) and symbols (SY) (Marcus, Marcinkiewicz, and Santorini 1993), but the corpus does not differentiate when these entities are used and mentioned. Even when a word that usually functions as a noun is mentioned, it is unclear what kind of noun (e.g., proper, common) it should be. In a way, it serves as a name, but it is not a *proper* name. Its referent (the word, as it appears) is peculiar when compared to the usual referents of nouns.

Complicating matters further, mentioned language often involves the production of linguistic entities in syntactic structures in which they usually do not appear. Letters, symbols, phonetics, and words that are rarely (if ever) *used* as nouns can be *mentioned*, imparting on them the qualities of nouns. These irregular appearances tend to lead parsers astray if they are trained on large labeled corpora.

```
(ROOT
  (S
    (NP (DT The) (NN word) (`` ``) (NN go) ('' ''))
    (VP (VBZ appears)
      (PP (IN on)
        (NP (DT the) (NN screen)))
      (PP (IN after)
        (NP (CD five) (NNS minutes))))
    (. .)))

(ROOT
  (S
    (NP (DT The) (NN word))
    (VP (VBP go)
      (SBAR
        (S
          (VP (VBZ appears)
            (PP (IN on)
              (NP (DT the) (NN screen)))
            (PP (IN after)
              (NP (CD five) (NNS minutes)))))))
    (. .)))
```

Figure 2.1: Parses of two stylistic permutations of the sentence "The word *go* appears on the screen after five minutes". The first contains quotation marks around *go*, while the second does not.

Consider, for example, the two parses of (24) in Figure 2.1, which were produced by the Stanford Parser (Klein and Manning 2003). The first parse is structurally correct, but the second parse does not have the apparent benefit of quotation marks around *go*. The output suggests that *go* is the main verb of the sentence, and this leads to an unusual (and unusable) construction of the sentence.

Although quotation marks are sometimes used to delimit mentioned language, they are a tenuous cue at best, as are all other stylistic cues. Quotation marks, bold text, and italic text are frequently indicators of mentioned language, but each of them has other common uses, such as emphasis and distancing. These cues are often unavailable in informal texts, and their spoken (approximate) analogues can be complicated to retrieve from speech recognition systems. Moreover, sometimes stylistic cues make little difference.

```
(ROOT
  (S
    (NP
      (NP (DT The) (JJ only) (NN word))
      (PP (IN on)
        (NP (DT the) (NN paper))))
    (VP (VBD was) (`` ``)
      (PP (IN before)))
    ('' '') (. .)))

(ROOT
  (S
    (NP
      (NP (DT The) (JJ only) (NN word))
      (PP (IN on)
        (NP (DT the) (NN paper))))
    (VP (VBD was)
      (ADVP (RB before)))
    (. .)))
```

Figure 2.2: Parses of two stylistic permutations of the sentence “The only word on the paper was ‘before’”. The first contains quotation marks around *go*, while the second does not.

Consider the two permutations of sentence (26) in Figure 2.2. The sentence in the first parse contains quotation marks around *before*, but neither parse assigns a reasonable label to the word.

Errors in parsing and tagging mentioned language have causes that extend beyond the parsing system used to generate the above examples. One problem is the irregularity of mentioned language: in any given corpus, a common word will appear in *use* far more often than in *mention*. A second problem is the “alienation” imparted upon mentioned language; the vocabulary and sentence constructions that usually surround a word are absent when the word itself is being discussed, leaving statistically trained parsers with little insight on what to do.

2.4.2 Conversational Systems

In contrast with written language, conversational language contains a higher frequency of metalanguage and mentioned language. One study (Anderson et al. 2004), using a subset of the British National Corpus, found that just over 10% of sentences in conversational English contain some form of metalanguage. Conversation contains a variety of challenges that interlocutors must overcome to understand each other and advance their respective goals. Channels between them are often “noisy” (in both the figurative and literal senses), leading to misunderstood or lost utterances. Conversational language is *informal* and often fraught with broken statements, restatements, sub-dialogues, and corrections. An interlocutor must model their dialogue partner’s knowledge state to understand what is appropriate to say next and to provide context for what has been said. New terms are sometimes introduced, and conversation history often must be referenced.

Superficially, it might seem bizarre that humans try to speak with each other at all. However, our facilities for recovering from perturbations in conversation are well-developed and require only nominal effort. These facilities are collectively termed *conversational adequacy* (Perlis, Purang, and Andersen 1998), and the ability to recognize the use-mention distinction is crucial to them. Interlocutors often utilize mentioned language or metalanguage to track dialogue state, clarify the meaning of terms, restate lost or misunderstood utterances, report others’ speech acts, and check dialogue partner comprehension. Moreover, our reliance on an understanding of the use-mention distinction runs deep, as it has been linked to the appearance-reality distinction in cognitive science (MJ Miller 1993). Without this understanding, a dialogue partner would likely be severely handicapped in their ability to participate in free-flowing, robust conversation.

To date, conversational agents (i.e., computerized dialogue partners) have had to function in spite of this handicap, often with frustrating results. Current conversational agents are susceptible to the *brittleness problem* in artificial intelligence (Anderson and Perlis 2005) when they are sidetracked by issues such as misunderstandings, misrecognitions, and user-realized errors (e.g., the user makes a wrong choice, knows it, and wishes to correct it). Those systems that do have recovery strategies for conversational perturbations tend to employ methods such as offering available choices again, repeating a question, or restarting a session. Such strategies come at a cost in time and user patience. This serves as motivation to develop conversational agents that, when appropriate, are capable of responding to problems in dialogue using the same familiar strategies that humans use when talking with each other.

2.4.3 Lexical Semantics

In order to understand language, humans must sometimes discuss language itself. Although much of our language learning happens implicitly, explicit statements about the meanings of words and phrases are essential for us as sources of direct and (relatively) unambiguous linguistic information. Situations occur when words and idioms must be introduced with mentioned language, either to highlight their introduction (alerting the audience of their importance) or to discuss semantics when doing so otherwise is difficult. After introduction, when the novelty of a new linguistic entity has diminished, mentioned language is often required to clarify or refine the audience's understanding of it.

Automated detection of mentioned language can make this condensed linguistic information available to language technologies as well, especially when such detection is applied to large corpora that contain many instances of the phenomenon. Such detection could, for instance, be used to prioritize (or place greater weight upon) the processing of sentences that contain explicit information about word referents. Explanations of idiomatic expressions could be identified, separated from the instances of their use, and given greater attention to determine their meanings. These techniques would be complimentary to existing approaches to corpora-based lexical semantics, in order to boost their performance.

2.4.4 Other Topics

The automated detection of mentioned language affects other topics of research, including:

- *Source Attribution*: When presented explicitly in discourse, quotation requires metalanguage and mention of language. An understanding of the structure of mentioned language will contribute to efforts to detect where in text sources are cited and what has been reproduced (e.g., delimiting sentences as mentioned without the aid of stylistic cues). This is especially true for the discussion of speech acts, when one person cites the utterances of another.
- *Sentiment Analysis*: Sentiments expressed in mentioned language are not always shared by the language user. For example, a person might talk about

an issue in a positive light while quoting others' arguments in order to refute them. Instances of mentioned language often provide additional information for a task, but for some tasks—such as sentiment analysis—those instances should be either reduced in value or discarded.

- *Natural Language Understanding*: A general problem remains open on how to represent statements about language so that intelligent inferences can be drawn from them. The flexibility with which humans discuss language and the variety of aspects of it that we refer to both pose a challenge to knowledge representation schemes.
- *Studies in Language Acquisition*: Studies have shown that humans employ the use-mention distinction in their efforts to understand language at an early age (Clark and Schaefer 1989). In research on second language acquisition, the value of the formal study of linguistic structure holds some controversy (Hu 2010), but the utility of informal metalanguage to discuss L2 words and concepts cannot be denied (e.g., "What does *llaves* mean?"; "You pronounce his name *row-OOL*"; "*Sum* is an irregular verb"). It is possible that certain metalinguistic strategies are more effective than others, and discovering this will require the ability to detect those strategies and accumulate information about them.

Although a full examination of these issues falls beyond the scope of this dissertation, they are noted here for the utility of this project to future research efforts.

2.5 Operationalizing the Definition

2.5.1 A Rubric for Mentioned Language

A human reader with some background in the use-mention distinction can often intuit the presence of mentioned language in a sentence, even in the absence of stylistic cues. However, to operationalize the concept and move toward corpus construction, it was necessary to create a rubric for labeling mentioned language. The rubric is based on substitution, and it may be applied, with caveats described in this section, to determine whether a linguistic entity is mentioned by the sentence in which it occurs.

Rubric: Suppose X is a linguistic entity in a sentence S. Construct sentence S' as follows: replace X in S with a phrase X' of the form "that [item]", where [item] is the appropriate term for X in the context of S (e.g., "letter", "symbol", "word", "name", "phrase", "sentence", etc.). X is an instance of mentioned language if, when assuming that X' refers to X, the meaning of S' is equivalent to the meaning of S.

Several examples will follow in Section 2.5.2 to demonstrate the rubric in action, and a rewrite of it as a series of instructions appears in Appendix B. Some minor adjustments in wording will be necessary for some sentences, and these will be shown. However, the handling of quotation marks, discussed below, must be addressed before continuing.

Quotation marks are frequently used as a stylistic cue for mentioned language, and they pose a slight problem to the rubric. This is because their inclusion or exclusion in the linguistic entity X can alter the meaning of the transformed sentence.

For instance, consider sentence (6) above when testing “Behind the eight”: if quotation marks are not included in X, then they surround “That phrase” in S’ and “Behind the eight” fails the rubric, since the meaning of the sentence has changed. (In fact, “That phrase” becomes mentioned language in S’.) If the quotation marks *are* included in X, then S’ does not contain them but “That phrase” refers to ““Behind the eight””, failing the rubric as well. Discarding the quotation marks altogether when applying the rubric might be sufficient in the general case, but it poses a problem for sentences like:

(41) The character “ tends to appear at the start of a quotation.

For the purposes of the rubric, quotation marks will be considered informal cues which aid a reader in detecting mentioned language. Style conventions may call for them, and in some cases they might be strictly necessary, but a competent language user possesses sufficient skill to properly discard or retain them as each instance requires. Similar reasoning can be applied to other stylistic cues, such as bold text and italic text, although those two cues have no literal representation in the string of characters that forms a sentence. To avoid further complications, examples in the following subsections will omit quotation marks or other stylistic cues.

Previous work on the use-mention distinction did not explicitly define the distinction or provide a procedure for verifying whether tokens qualified as mentioned language. The rubric has some distinct advantages over those previous efforts; the advantages will become apparent in the literature review (Chapter 3, which contains many more citations) but are appropriate to summarize here:

1. *The rubric does not depend on the presence of quotation marks*. Many previous theories of the use-mention distinction require quotation marks to play an integral role in mentioning language (Quine 1940; Davidson 1979; Tarski 1933; García-Carpintero 2004)[6]. In many cases, competent language users are able to recognize the use-mention distinction without the aid of quotation marks. (The reader may verify this by reading the example sentences in the next section without reading their accompanying paragraphs.) Quotation marks (along with most proper punctuation) are often omitted in informal contexts, and even in formal writing other cues lacking a literal presence (particularly bold and italic text) sometimes take their place. The rubric acknowledges the flexibility of the human recognition of the use-mention distinction, while the previous work does not.
2. *The rubric provides an explicit procedure for identifying mentioned language*. Earlier efforts tend to assume that examples of the use-mention distinction are obvious, and they work from that assumption to an explanation of the semantics of mentioned language (Cappelen and Lepore 1997; Maier 2007). The rubric provides a mechanism for verifying purported examples, a desirable alternative to accepting them at face value.
3. *The rubric exploits the relationship between mentioned language and its sentential context.* Even when quotation marks are consistently used, it is the

[6] In fact, previous literature surveys (Saka 1998; Anderson et al. 2002) have shown that *nearly all* of the proposed theories of the use-mention distinction have required quotation marks to delimit mentioned language . This practice is so widespread that the literature often uses the term *quotation* to refer to mentioned language, causing confusion over the meaning of the term, as it also refers to the reproduction of language from another source. Reproduction of language is a function of mentioned language but certainly not the only function.

sentential context of mentioned language that selects the relevant aspect of a mentioned token (e.g., spelling, pronunciation, meaning, etc.); without this, the meaning of an instance of mentioned language is uncertain. Notably, Saka's ostention theory (1998) recognizes the value of sentential context but does not have the advantage of item 2 above.

2.5.2 Examples of Rubric Usage

Some examples will illustrate how the rubric covers the varieties of mentioned language. For instance, consider

(42) Fancy automobiles are called luxury cars.

where the phrase "luxury cars" is under consideration. Choosing "that phrase" as a replacement, the sentence becomes

(43) Fancy automobiles are called that phrase.

where "that phrase" is understood to refer to "luxury cars". While there might be contextual ramifications (for instance, if a language user specifically wants to utter the phrase "luxury cars"), the reader can verify that the meaning of the sentence is essentially unchanged, and "luxury cars" passes the rubric. In contrast, consider testing the phrase "Fancy automobiles" in (42). The substitution test and a charitable adjustment to the verb phrase result in

(44) That phrase is called luxury cars.

where “That phrase” is understood to refer to “Fancy automobiles”. It is plausible (albeit odd) that a speaker might wish to assign the name “luxury cars” to the phrase “Fancy automobiles”, but it is clearly not the intent of the original sentence. The phrase “Fancy automobiles” in (42) fails the rubric.

This rubric requires some adjustment when the sentence already explicitly refers to X as a word, phrase, or other appropriate entity, such as in (2), (7), or (9) above. In such cases it may be appropriate to omit the linguistic entity under consideration without substituting, such as this alteration to (9):

(45) That French word refers to a feline animal.

where “That word” is understood to refer to “chat”. Explicit discussion of word categories requires similar omission without substitution, as in testing “help” in (46) with the rubric substitution shown in (47) below:

(46) The verb help has several different senses.

(47) That verb has several different senses.

In (47), “that verb” is assumed to refer to “help”; this does not change the meaning of the sentence, and “help” passes.

The rubric also accounts for explicit discussion of pronunciation, as shown by testing “canz” in (48) below with its substitution equivalent (49):

(48) Australians say the city name as canz.

(49) Australians say the city name as that pronunciation.

In (49), “that pronunciation” is assumed to refer to the pronunciation represented by *canz*. Although the written medium requires an additional level of interpretation—to distinguish the sequence of letters *canz* from the pronunciation of *canz*—the reader should infer this from the substitution phrase. “canz” thus passes the rubric.

Discussion of characters and spelling is similarly covered, as shown by (50) (testing “ie”) and its substitution equivalent (51) below:

(50) Garcia spelled his name with ie back then.

(51) Garcia spelled his name with that character string back then.

where in (51) “that character string” is assumed to refer to “ie”. This does not change the meaning of the sentence, and “ie” passes the rubric. However, it might be tempting to test “Garcia” in (50) as well, and the substitution equivalent of the sentence is (52) below:

(52) That character string spelled his name with ie back then.

Sentence (50) asserts that Garcia is responsible for the said spelling, and sentence (52) asserts that a character string is responsible for the spelling. Since these are very different meanings, “Garcia” fails the rubric and is not mentioned language. Similarly, terms that simply *refer* to characters or symbols do not pass the rubric. Consider the test of “The symbol for infinity” in (53) below, shown with its substitution equivalent in (54):

(53) The symbol for infinity is a lemniscate.

(54) That phrase is a lemniscate.

In (54), “That phrase” is understood to refer to “The symbol for infinity”. Sentence (53) asserts that the symbol for infinity is a lemniscate. Sentence (54) asserts that “The symbol for infinity”—*that phrase*—is a lemniscate. Since these meanings are different, “The symbol for infinity” fails the rubric.

The rubric also shows that sentences with almost identical wording can differ on whether they use or mention the same word or phrase, as in (55) and (56) below (testing “Spain” in both):

(55) Spain is the name of a European country.

(56) Spain is a European country.

Their substitution equivalents are, respectively:

(57) That name is the name of a European country.

(58) That name is a European country.

where in (57) and (58) “that name” refers to “Spain”. Sentence (57) effectively asserts that “Spain” is the name of a European country, which does not change the meaning of the sentence; thus, “Spain” in (55) passes the rubric. However, (58) asserts that “Spain”—the name itself—is a European country. Since this changes the meaning of a sentence, “Spain” in (56) fails the rubric.

While many other permutations exist that require minor adjustments in wording, such untamable variation is inherent in natural language, and the spirit of the rubric will be sufficient for the studies in subsequent chapters.

2.5.3 The Definition-Rubric Relationship

The main intent of the rubric is to validate all and only the linguistic entities that qualify for the definition when the T in the definition draws attention to the *type* of T[7]. This partial equivalence is illustrated with the conjunction of two claims.

Claim #1: if a token T is produced in a sentence to draw attention to a property of the type of T, then the deictic substitution in the rubric does not alter the truth conditions[8] of the sentence (thus satisfying the rubric). This is because the properties of a type are not altered by the substitution in the sentence. In sentence (59) below

(59) “Cat” has three letters.

“Cat” is mentioned to draw attention to the spelling of the type of the word. In (60) below this is preserved:

(60) That word has three letters.

assuming, as the rubric requires, “That word” is understood to refer to “Cat”. Properties of the type of “Cat” remain intact independent of its presence or non-presence in the sentence.

Claim #2 is the converse of the first: if applying the rubric substitution to a token T in a sentence does not alter the truth conditions of the sentence (thus satisfying the rubric), then the original sentence draws attention to a property of the

[7] The rubric also covers some (but not all) instances of mentioned language where T is produced in a sentence to draw attention to the *token* of T. This will be discussed further in Section 2.5.4.

[8] For present purposes. *truth conditions* will be considered equivalent to *meaning*. However, a substantial body of work exists on articulating the relationship between these two, as surveyed by Lynch (2001) and Kirkham (1995).

type of T, satisfying the definition. This is because the deictic phrase must have a relationship with the content of the sentence, and that relationship must match that of the original token (T) that it displaced. Consider (61) below and (62), with the rubric transformation on “cheese”:

(61) Cheese is derived from a word in Old English.

(62) That word is derived from a word in Old English.

These two sentences are equivalent in meaning because both of them refer to a property of “Cheese”, regardless of the need to resolve the referent of “That word” in (62). By referring to a property of “Cheese”, (61) satisfies the definition. Moreover, the rubric indicates that “Cheese” in (63) below (transformed in (64)) is not mentioned language:

(63) Cheese is derived from milk.

(64) That word is derived from milk.

Sentence (63) refers to a property of cheese and (64) refers to a property of an unknown word indicated by “That word”. Sentence (63) thus does not satisfy the definition, either.

Together, Claim #1 and Claim #2 above should be sufficient to establish that the rubric as plays the same formal role as the definition in a wide variety of cases.

2.5.4 SRT Sentences

Sentences that mention T to draw attention to its token *and* invoke self-reference to do so will not be handled by the rubric. (For brevity these will be called

"SRT sentences"—Self-Reference to a Token.) For example, consider testing SRT sentence "Cat" in (65) below, with the substitution result shown as (66):

(65) Cat is the first word in this sentence.

(66) That is the first word in this sentence.

whereas "That" in (66) is understood to refer to "Cat". Although this sentence satisfies the definition of mentioned language, the substitution changes its meaning. Sentence (65) claims that "Cat" is the first word in sentence (65), which is true. Sentence (66) claims that "Cat" is the first word in (66), which is false. Notably, token-mention alone does not pose a problem for the rubric; consider (67) and (68) below, testing the rubric on "cat":

(67) The token cat has three letters.

(68) That token has three letters.

whereas "That token" is understood to refer to a tokenization of "cat". These sentences share the same meaning, even though one contains the token "cat" while the other makes a deictic reference to a tokenization of "cat". Sentence (67) thus passes the rubric. Also, self-reference alone does not trouble the rubric; consider (69) and (70) below, testing the rubric on the second appearance of "which":

(69) This sentence, which contains the word which, has two commas.

(70) This sentence, which contains that word, has two commas.

whereas "that word" is assumed to refer to "which". The reader may verify that the meaning of (69) does not differ from (70).

Since SRT sentences can only illuminate properties of a token that involve the token's relationship with the sentence, these sentences appear to be of limited or no practical value for inclusion in the corpora created by this project. Excluding them from the rubric was deemed fair since the goal of creating it was practical: to operationalize the definition and simplify manual labeling of mentioned language. Still, for their peculiarity, SRT sentences may deserve further examination in future theoretical studies.

2.6 Conclusions

This chapter introduced the use-mention distinction, with particular attention to mentioned language, since its detection will be the focus of this dissertation. A definition for the phenomenon was presented, along with a rubric for detection by hand. An illustrative list of the varieties of the phenomenon was created, along with explanations of how each variety will or will not be handled by this study. The benefits of detecting mentioned language were discussed for parsing, conversational systems, and other active areas of research. Finally, justification was given for accepting the rubric as properly playing the same formal role as the definition in a wide variety of cases.

Chapter 3: Related Work

3.1 Introduction

The use-mention distinction has a long and varied history of study in theoretical linguistics and the philosophy of language, but it has received little examination in computational fields. There have been many attempts to explain how language users are able to interpret the distinction, what precisely mentioned language represents, and what insight the distinction should give us into other language phenomena. These theories tend to agree on little, and the mere existence of the use-mention distinction is a notable concordance, though even that existence is occasionally questioned (Christensen 1967).

Section 3.2 divides theories of the use-mention distinction into three categories. *Exclusive theories* are those that divide used language and mentioned language into two separate phenomena with little or no overlap. *Non-exclusive theories* hold that use and mention stem from a set of referents embedded in language (called *ostentions*) that allow use and mention to occur simultaneously. *Applicative theories* are not concerned with the origins of the distinction so much as phenomena in language that the distinction supports and allows language users to exhibit. Sections 3.3 and 3.4 will address related work in computational fields, and 3.5 will discuss the relationship between the related work and the present study.

3.2 Theories of the Use-Mention Distinction

3.2.1 Exclusive Theories

Theories that separate use and mention dominated the discussion of the distinction through much of the twentieth century. These theories often refer to *quotation* rather that *use-mention*, owing to the prevalent stylistic cues (i.e., quotation marks) and perhaps terminological brevity. However, the present study does not treat those terms as synonyms, and this discussion will use them as defined in Chapter 2. The theories can be divided into five groups, examined in this subsection with inspiration from literature surveys by Saka (1998) and Anderson, et al. (2002).

Name theories consider an instance of mentioned language to be a name for its referent. Tarski was the original purveyor of this theory, explaining that

> Quotation-mark names may be treated like single words of a language…the single constituents of these names…fulfill the same function as the letters and complexes of successive letters in single words. Hence they can possess no independent meaning. (Tarski 1933)

Quine's analysis was similar, declaring that "each whole quotation must be regarded as a single word or sign…The meaning of the whole does not depend on the meanings of the constituent words" (1940). This theory might have been suitable for single symbols in mathematical logic, but in the realm of natural language it is insufficient. In a sentence like

(1) Lincoln's title was "President of the United States".

the reader cannot help but infer meaning into the phrase *President of the United States*: it reveals (assuming no deception) that Lincoln was the president of the United States. The name theory is perhaps sufficient for cases like

(2) Lincoln’s title was “qarabzug”.

where the mentioned phrase implies no conventional referent and it is introduced as a novel term. However, such cases are only a fraction of all instances of mentioned language. This theory also accounts for only a single purpose of mentioned language, the presentation of names, which makes it incompatible with the motivations of the present study.

Description theories provide a slight improvement over name theories by reducing the opacity of mentioned language. An instance of mentioned language is interpreted as a series of tokens which describe the referent phrase. Some descriptive theories hold that the tokens are letters or phonemes (Quine 1940; Richard 1986; Tarski 1933), while others consider words to be the tokens (Geach 1950). For example, the quoted text in (1) is a description of the phrase *President of the United States* by virtue of being this sequence of words: *President, of, the, United, States*. Although description theories acknowledge some amount of structure in mentioned language, they still fail to account for the audience’s semantic aptitude when faced with the phenomenon. This makes a description theory no more suitable than a name theory for present purposes.

Demonstrative theories propose that quotation marks (and other stylistic cues, presumably) “point to” the representation of the quoted material (Davidson 1968; Goldstein 1984). Under this theory, (1) could be interpreted as

(3) President of the United States. That was Lincoln’s title.

with the implicit assumption that *That* refers to the text preceding it. This theory does not preclude the existence of semantic depth for mentioned language, which is an improvement over name and description theories. It also provides a basis for transformation (as from (1) to (3)) to reinterpret mentioned language without stylistic cues; in a way, this simplifies mentioned language. However, the theory is *dependent* upon those stylistic cues, which is problematic because they are often unnecessary for an audience to recognize and interpret mentioned language. Nested use of stylistic cues is also a problem, since it is unclear how to interpret such occurrences. The transformation advocated by demonstrative theories provided some inspiration the rubric presented in Subsection 2.5, but beyond that it will not be utilized.

Identity theories (Reimer 1996; Washington 1992) propose that text between quotation marks (and inside other stylistic cues, presumably) refers to itself. This contrasts sharply with demonstrative theories, which consider stylistic cues to be pointing devices; for an identity theory, the cues are semantically inert. It is unclear then why we use them at all, since the capacity to recognize mentioned language must come from other means, at least partially—a notion that, initially, seems appropriate. However, this “semantic inertness” leads to conclusions that clash with common conventions on the use of stylistic cues. At least intuitively, *“President of the United States”* (quotes italicized intentionally) and *President of the United States* refer to the title *President of the United States* and the person President of the United States, respectively. It would be difficult for an identity theory to conclude that they refer to different things, whether the title or the person. Moreover, such a theory does little to explain the use-mention distinction, and while that is not strictly a fault, it is a highly desirable property.

Finally, *syntactic theories* reflect properties of more than one of the previous categories. Partee (1973) proposes that *word quotation* and *sentence quotation* ought to be examined separately. In sentence quotation, she proposes that text between quotation marks is a demonstration of its surface structure, which is its sole contribution to the meaning of the containing sentence. In a very limited set of cases, such as (4) below, that seems tenable:

(4) “Suzy likes dark chocolate” is a sentence.

The only aspect of *Suzy likes dark chocolate* that (4) refers to is its satisfaction of the (chiefly syntactic) properties of a sentence. However, it is easy to conjure examples that mention sentences for aspects other than their syntax, such as meaning or attribution:

(5) “Elizabeth likes milk chocolate” means Elizabeth likes milk chocolate.

(6) The vet said “Frisky is a happy cat.”

Cram (1978) argues that instances of text between quotation marks fulfill the requirements of noun phrases and should be treated as such. While this treatment is frequently consistent with the functions of mentioned language in a sentence, it is sometimes inconsistent (as in partial quotation) and fails to account for any internal structure that mentioned language may possess. Neither Partee nor Cram account for the occurrence and comprehensibility of quotation or mentioned language without quotation marks.

3.2.2 Non-Exclusive Theories

To their credit, the exclusive theories of the use-mention distinction can be satisfactory for formal or mathematical languages. Tarski and Quine were principally concerned with mathematical logic, and the dependence of their theories (as well as others' theories) upon quotation marks is perhaps an artifact of the mathematical pedigree. This rigor comes at a price, however, as the exclusive theories cannot cope with the fluidity of natural language. Language-mention without stylistic cues and partial quotation both thwart the exclusive theories, but they face an even more fundamental problem. Most instances of mentioned language (as in (1) above) invoke some audience understanding and thus language-use, even if it is not primary to the purpose (as in (4)). One could easily argue that most quotation contains a mixture of use and mention, with a few exceptions like (2).

Some recent theories of the use-mention distinction *do* allow for use and mention to occur simultaneously. García-Carpintero (2004) and Saka (1998) both discuss the use-mention distinction in terms of *ostentions* of language.[9] García-Carpintero observes that

> [W]e do not merely refer with quotations to expression-types, but also to other entities related in some way to the relevant token we use: features exhibited by the token distinct from those constituting its linguistic type, features exhibited by other tokens of the same type but not by the one actually used (as when, by using a *graphic* token, we

[9] Saka and García-Carpintero both claim that their counterpart's theory is not truly ostensive (García-Carpintero 2004; Saka 2003). At the present level of discussion, it will not be necessary to explore their reasons for disagreement.

> refer to its phonetic type), or even other related tokens[.] (García-Carpintero 2004)

Similarly, Saka writes:

> [E]very expression token (e.g. this particular inscription: cat) ambiguously or indeterminately refers to itself and to various items associated with it (including the inscription-*type* “cat”, the pronunciation /kæt/, the concept CAT, and the extension of cats). Quote marks…help to disambiguate the intended reference, although they are usually neither necessary nor sufficient for doing so. (Saka 1998)

He later applies the term *ostention* to these features or items associated with language tokens. Any token is capable of invoking multiple ostentions at once in a reader’s mind; the reader discriminates among them using a variety of cues. Stylistic cues (or their counterpart cues in spoken language) aid an audience in choosing ostentions for each token in an utterance; writers may use them to follow conventional practices or to precisely delimit mentioned language. They also may omit the stylistic cues in some circumstances, if such cues are unnecessary or impossible to use.[10]

Saka contends that there are “at least” five ostentions for recognized words in a language, listed here for *cat*:

[10] This is the reason why this dissertation uses the term *language-mention* to refer to the phenomenon that most of the literature calls *quotation*. It is quite possible—and entirely acceptable—for some instances of language-mention to appear without quotation marks or other stylistic cues. It would be awkward to use the term *quotation* to refer to a sequence of tokens that is *not* surrounded by quotation marks, especially since stylistic cues will feature prominently in the detection of mentioned language, presented in later chapters. The term is thus redirected to one of its other common meanings, the reproduction of language from another source, as explained in Chapter 2.

(a) orthographic form: cat

(b) phonic form: /kæt/

(c) lexical entry: <cat, /kaet/, count noun, CAT>

(d) intension: [the concept of] CAT

(e) extension: {x: x a cat}

For a token or sequence of tokens, exposure to either (a) or (b) in some context initiates a disambiguation process for any competent speaker of English. In a successful case of language-use, the audience understands that (d) or (e) is being ostended. In a case of language-mention, the audience understands that (a), (b), (c) or another item associated with the token(s) (e.g., grammaticality, truth value, quotative properties) is being ostended. Crucially, a reader can interpret a sequence of tokens as both use *and* mention. This explains how all example sentences in this chapter except (2) contain mentioned language that the reader inevitably interprets in some non-mention capacity as well.

The ostensive theory can be applied to some phenomenon not traditionally thought of as language-mention, such as irony and sarcasm (Saka 2003). Such phenomena "draw attention" to language while being used, even if the effect is less dramatic than it is for statements directly about language. One might even argue that *all* language is a mixture of use and mention, even if mention is often present only in a tenuous sense, since choices in language use (e.g., choosing one word over another, choosing to produce language in some context instead of remaining silent) can convey paralinguistic information.

3.2.3 Applicative Theories

Cappelen and Lepore (1997) posit four categories of language-mention, introduced briefly in Chapter 2 and explained here using the original examples from their articles. Assuming, first, that Alice utters (7) below

(7) Life is difficult to understand.

direct quotation mentions her full statement:

(8) Alice said “Life is difficult to understand”

Indirect quotation (termed *paraphrase* in this dissertation) reports what she said while not necessarily (though possibly) using her original words:

(9) Alice said that life is difficult to understand.

Mixed quotation reports what she said while only necessarily using some of her words:

(10) Alice said that life “is difficult to understand”.

Finally, *pure quotation* (termed *pure mention* presently) is produced by a language user simply to talk about linguistic expressions:

(11) “Life is difficult to understand” is a sentence.

Cappelen and Lepore observe that there are semantic commonalities between all four of these categories, and any treatment of one should explain the others as well. They embark on a project to represent them using first order predicate logic, while also explaining the relationships between each of the categories and Davidson’s (1979) demonstrative theory of quotation.

Maier (2007) investigates three categories of quotation which overlap those of Cappelen and Lepore but receive different terminology. *Direct discourse* faithfully reproduces an utterance in its entirety, including any errors. This is illustrated with a quote from George W. Bush[11]:

(12) [Bush:] “I’ve, I’ve got a eckullectic reading list.”

Indirect discourse (introduced as *paraphrase* in Chapter 2) eschews exact wording to focus instead on what is expressed:

(13) Bush said that he has an eclectic reading list.

In *mixed quotation*, a particular phrase is quoted verbatim, but the rest is reported indirectly[12]:

(14) Bush said that he has an “ecelectic” reading list.

Maier observes that quotation poses several challenges to formalization. For instance, the acceptability of a report is not affected by errors in quoted language, as in (1). On the other hand, indexicals require adjustment: “I” in the direct discourse of (12) and “he” in the indirect discourse of (14) both refer to the same person. Maier also taps

[11] Maier’s original URL citation for this quote is no longer valid, but an alternate exists: `http://www.youtube.com/watch?v=eKiWWi8rdJQ&feature=related`

[12] Maier’s original citation (still functional): `http://www.thecarpetbaggerreport.com/archives/8339.html`

Davidson’s demonstrative theory of quotation (1979) to help explain the phenomenon.

3.3 Natural Language Processing

Occurrences of metalanguage have received relatively little attention in corpus-based natural language studies, which makes notable the contribution of Anderson et al. (2004). Their paper describes the process of developing an annotation scheme for metalanguage in conversational English and its application by hand to a subset of the British National Corpus (BNC). Their annotation scheme uses five major categories, listed here with their examples:

- Track Dialog (TD): “Which particular section of the conversation are we talking about?”
- Speaker Meaning (SM): [first speaker] “I had a right argument over that.”; [second speaker] “Who did, them two?”; [first speaker] “No, me and Laura did.”
- Language Meaning (LM): “Yes, as well, binge, binge, not ‘bilge’.”
- Determine Truth (DT): [first speaker] “I’d rather be working.”; [second speaker] “Oh, God. You don’t really mean it?”
- Speech Acts (SA): “Yeah, we remember when you shouted ‘here she comes’.”

The study determined that 10.94% of the sentences in the selected BNC subset contained at least one instance of metalanguage. The most common annotation by far was SA (72% of the occurrences), followed by SM (15.13%), TD (7.66%), LM (3.83%), and DT (1.10%). (A remaining 0.18% of the occurrences were unclassified due to human disagreement or unsuitability of the five categories.) Also studied was the possibility that certain words occur in higher frequency in sentences with metalanguage. The report identifies ten words with positive predictive values (PPVs) for metalanguage greater than 0.5 (though more may exist), the highest three being “said”, “pardon” (though with few occurrences), and “say/s/ing”. While these results are encouraging, the highest PPV is 0.84 and seven of them fall below 0.7, suggesting the need to find additional metalanguage clues.

Another area of study related to metalanguage is discourse. Lin et al. (2009) describe the creation of the Penn Discourse TreeBank (PDTB), a corpus built on top of the Penn Treebank (Marcus, Marcinkiewicz, and Santorini 1993) and PropBank (Palmer, Gildea, and Kingsbury 2005). To create the corpus, readers manually annotated discourse connectives and their arguments, although the levels of agreement between readers varied widely depending on the connectives. The group also published a study (Dinesh et al. 2005) comparing the annotations of discourse arguments in the PDTB with annotations (from the original Penn Treebank) of syntactic structure in the same text. They found significant differences between the two, and showed how those differences were due “in large part” to the *attribution* of discourse arguments and the connectives themselves to speakers. Discourse structure is not always a factor in language-mention, but exploitable relationships exist between them, such as the presence of language indicating source attribution.

Another group (Riloff, Wiebe, and Phillips 2005) studied a similar topic, subjectivity classification, with the goal of filtering out subjective statements to improve the accuracy of information extraction systems. To do this, they developed a method for identifying subjective language (e.g., opinions, metaphor, hyperbole) in text at the sentence level. They start with a rule-based classifier that consults a list of fairly certain subjectivity clues, which is then applied to an unlabeled corpus to separate subjective and objective text. This data is then fed as training to a Naïve Bayes classifier, which examines a greater variety of features than the original rule-based classifier. Several experiments showed that the pre-filtering did improve the performance of information extraction on standard problems, though some fine-tuning was required: source attribution, although a sign of subjectivity, was often a strong clue that a sentence contained facts worthy of extraction.

Finally, English Wikipedia will be used as a text corpus in experiments described in this dissertation. The emerging utility of Wikipedia as a corpus is well-documented in the literature. A few related uses of Wikipedia include named entity recognition (Balasuriya et al. 2009), syntactic parsing (Honnibal, Nothman, and Curran 2009), and lexical semantics (Zesch, Müller, and Gurevych 2008), among many others. Ytrestøl et al. (2009) previously noted the relationship between stylistic cues in Wikipedia and the use-mention distinction, though this observation was incidental to their focus on the automatic extraction of sub-domains of articles.

3.4 Commonsense Reasoning About Language

The use-mention distinction has overlapping ramifications for research in natural language, conversational agents, and reasoning. This section will explore related work in the overlap, concerning commonsense reasoning about language.

Much research has focused on crossing the boundary between natural language and formal representations of knowledge. ConceptNet (Havasi, Speer, and Alonso 2007) is a large (250K element) semantic network of commonsense knowledge. ConceptNet represents knowledge using natural language fragments, instead of formal logic, to mirror the defeasibility and context-sensitivity of human commonsense knowledge. For example, an excerpt of ConceptNet's network might show nodes "person" and "feel jittery" connected by the edge "do not want", representing the fact that a person does not want to feel jittery. This representation method shares some of the disadvantages of natural language, such as redundancy when the same concepts are represented multiple times by different words. Liu and Singh (2004) provide examples of activities possible with ConceptNet, such as context finding, inference chanining, and classifying conceptual knowledge.

Anderson et al. (2002) discuss the need for conversational systems to be capable of meta-dialogue. They coin the term *conversational adequacy* to describe the ability to engage in flexible, free-ranging conversation, and they explain how meta-dialogue and the use-mention distinction are central to that ability. Among other uses, these mechanisms allow dialog partners to establish grounding by referring to the conversation, correcting misunderstood communication, checking on dialogue partner comprehension, and introducing new terms to the discourse. An earlier paper (Perlis, Purang, and Andersen 1998) also discusses conversational adequacy in greater detail, proposing that "meta-dialog and meta-reasoning are, in some sense, both necessary and sufficient for communication". The authors describe this thesis in detail:

(i) *Sufficiency*: as long as there is at least a *weak* ability in the object capacities (inference, learning, and language) then effective conversation can proceed if there is a *strong* miscommunication competence.

(ii) *Necessity*: no matter how strong the object capacities, effective conversation cannot proceed if there is not a strong miscommunication competence.

The ability to recognize and reason about language-mention underlies conversational adequacy, since it is our facility for recognizing when dialogue is being discussed.

Basic strategies for conversational adequacy have been implemented in the previously-discussed ALFRED (Josyula et al. 2007) dialog agent. ALFRED (Active Logic for Reason-Enhanced Dialog) is described as a "universal interfacing agent", which a user converses with via a subset of natural language in order to control a variety of task-oriented domains. ALFRED reasons in time about dialog in a manner meant to resemble human cognition, and it is capable of some forms of meta-dialogue to recover from anomalies in conversation. For instance, when the user employs a word unknown to ALFRED in an utterance, the system asks for clarification (generally in the form of a synonym), enters the clarification into its knowledge base, and then returns to processing the original utterance.

ALFRED is one of a few recent systems designed for commonsense reasoning about language. Another is SNePS (S Shapiro et al. 2007), which employs a modified form of first-order predicate logic to model commonsense reasoning and natural language understanding. The modifications are designed to more accurately and elegantly mimic statements in natural language, using four types of expressions: propositions, rules, acts, and individuals. While both ALFRED and SNePS use predicate logic to represent knowledge, the focus of the former is to demonstrate aspects of conversational adequacy, and the focus of the latter is to accurately represent complex statements from natural language.

Finally, Anderson and Perlis (2005) propose a highly generalized approach to handling anomalies, not only for conversational systems but also for any autonomous systems that must be tolerant of perturbations. They propose the term *brittleness* to denote the common problem of an autonomous system failing when faced with a situation even slightly outside of its original programming. A human, when faced with an unforeseen perturbation to a plan, notes the nature of the problem, assesses the reason behind it, and guides a solution into place. This process of reasoning was termed *the metacognitive loop*, and it was implemented in ALFRED (described above) and a computer agent that plays Bolo, a multiplayer video game (MD Schmill et al. 2007).

3.5 Discussion

Section 3.2 presented several theories of the use-mention distinction, many with strong influences from mathematical logic and philosophy. The lack of practical attention to the distinction in natural language is perhaps the reason why very little is agreed upon for its mechanics and terminology. By studying occurrences of patterns of language-mention “in the wild”, this dissertation will resolve some of the uncertainty surrounding the phenomenon. Still, the previous theoretical work will provide a valuable framework for this effort. Given the available alternatives, the ostensive theory of the use-mention distinction will be the dominant paradigm for this study, since it has these advantages:

- It describes a flexible relationship between mentioned language and stylistic cues. Other theories assume implicitly that stylistic cues are always present, or they assign them to a crucial, always-present semantic task. Under the ostensive theory, stylistic cues are merely a disambiguation aid to help the audience choose between relevant ostentions. The cues are necessary in some cases of mentioned language, obligatory (but optional) in others, and completely unnecessary in still others.

- It admits the different reasons why humans exhibit mentioned language. Quotation, in the sense of reporting an utterance, is only one of these purposes, as is “pure” mention without semantics. Many different properties of a token or type can be highlighted by language-mention, and Saka’s list of ostentions is an intuitive (though incomplete) list of these properties.
- It accounts for simultaneous use and mention. Lost in some other theories of the distinction is the fact that humans do not suspend their language understanding facilities when language-mention occurs. The ability for humans to process multiple ostentions at once is not just serendipitous but also sometimes required, as shown in Category K (mixed referent in sentence) from Chapter 2.

Meanwhile, little previous work was available on the use-mention distinction in computational fields, such as natural language processing and commonsense reasoning. A paper by Anderson et al. (2002) drew attention to this lack, and later a metalanguage corpus (Anderson et al. 2004) became this study’s closest ancestor, but it produced only a survey of categories and some observations on their distribution. It will be incumbent upon the present study to delve deeper into patterns that can be used for reliable detection of mentioned language.

Chapter 4: Creation of a Robust Corpus of Mentioned Language

4.1 Introduction

To study the use-mention distinction, it was necessary to gather a large, diverse sample of occurrences of mentioned language. Although "laboratory examples" (such as many of those in this dissertation) begin to illustrate variations in the phenomenon, instances gathered from a large body of language provide a better picture of how humans exhibit the phenomenon. This chapter describes the process of building a series of three corpora of mentioned language, containing sentences from English Wikipedia. The corpora were built using progressively more sophisticated methods, with each construction procedure refined using lessons learned from previous results. The *Pilot Corpus* was built to verify that mentioned language could be gathered from Wikipedia using stylistic cues, and to gather a set of "mention words" to enable better candidate filtering. The *Combined Cues Corpus* was built to determine whether the combination of lexical and stylistic cues would produce a rich mixture of mention candidates. Finally, the *Enhanced Cues Corpus* was built to create a robust final corpus to study automated identification of mentioned language.

4.2 The Pilot Study

This section describes a pilot effort to build a small corpus of instances of mentioned language. The effort was motivated by three goals:

1) to verify a hypothesis that stylistic cues (i.e., bold text, italic text, text between quotation marks) are an effective heuristic for gathering instances of mentioned language;

2) to begin examining patterns of the phenomenon that might later be used to identify it in absence of stylistic cues; and
3) to determine the applicability of the ostensive theory of the use-mention distinction in practice.

Section 4.2.1 discusses the choice of *Wikipedia* as a source of text for corpus building. Section 4.2.2 details the creation of the pilot corpus, and 4.2.3 describes the composition of the corpus that was created and makes some observations on it.

4.2.1 Rationale for Choosing Wikipedia

The article set of *English Wikipedia*[13] was chosen as a source for text, from which instances were mined using automated and manual methods. This section explains the reasons why Wikipedia was selected and how some practical considerations were met.

Wikipedia is a particularly suitable source for collecting instances of mentioned language. Listed here are four factors that led to its selection for this study:

1) *Wikipedia is written to introduce a wide variety of concepts to the reader*. At the time of this dissertation draft, Wikipedia contains approximately 3.5 million articles. These articles are written informatively and they generally assume that the reader is unfamiliar with the topics they discuss. New names and words are frequently introduced, often explicitly and in a manner that invokes language-mention.
2) *Stylistic cues that are sometimes used to delimit mentioned language are present in article text*. Wikipedia contributors generally use quote marks, italic

[13] Described in detail at `http://en.wikipedia.org/wiki/English_Wikipedia`.

text, or bold text to “highlight” where language is mentioned. This convention is stated in Wikipedia’s own style manual[14], though it is unclear whether most contributors read it there or follow it out of habit. Although these cues are used for other activities in addition to language-mention, they provide a starting point for automatic extraction of the phenomenon.

3) *Wikipedia is collaboratively written*. Since any registered user can create, contribute to, or edit articles, Wikipedia text reflects the language habits of a large sample of English writers. It is unclear how much variation exists between writers on how to mention language, so this large sample is desirable.
4) *Wikipedia is freely available*. Language-learning materials (particularly textbooks) were also considered, but issues of legality and electronic availability were deemed obstacles. Wikipedia’s licensing of article text is compatible with the goals of this project[15], and downloading articles *en masse* is uncomplicated. Moreover, the markup code for Wikipedia articles is easy to access and interpret.

Choosing Wikipedia for this project introduced some limitations as well. Three of them are listed below, with responses to each.

1) *Limitation: Wikipedia article text is written in (relatively) formal language*. Articles are written to inform, and this purpose (combined with Wikipedia’s own internal language culture) leads to a certain encyclopedic style and

[14] Available at http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_text_formatting.

[15] Wikipedia’s article text is available for reuse under the Creative Commons Attribution-ShareAlike 3.0 Unported License and the GNU Free Documentation License. For details, see https://secure.wikimedia.org/wikipedia/en/wiki/Wikipedia:Copyrights#Re-use_of_text.

disposition. Less formal domains that also contain language-mention include personal blog entries, forum posts, and instant messages.

Response: The lack of previous work in language-mention detection was an influence in the choice of a relatively orderly source of text; it did not make sense to begin with difficult material. The unique advantages of Wikipedia listed previously, combined with the project hypothesis (that lexical and syntactic cues are sufficient to detect most language-mention), led to the selection. This effort should produce a relatively "clean" set of mention cues, which are expected to be applicable to other domains as well with adjustments only for their informality.

2) *Limitation: Stylistic cues are imprecise indicators of mentioned language*. They serve other additional roles in language, such as distancing ("scare quotes"), emphasis, and *use* of names and titles. Wikipedia articles lack consistent editing, so some instances of mentioned language might not be highlighted by the stylistic cues.

   Response: Stylistic cues make the gathering of instances of mentioned language a tractable problem. Without them, it would be necessary for a human reader to examine *all* text in a corpus to gather candidate instances of the phenomenon and then manually delimit them. (Using the cues as a pre-filter, some reading will still be required, but much less.) Since Wikipedia is collaboratively written, systematic failures to highlight mentioned language are not expected to occur over a large sample of articles.

1) *Limitation: The applicability of the findings using Wikipedia, a specific source of text with its own nuances, is unknown.* Even if Wikipedia is uniquely suitable for gathering mentioned language, this is not a guarantee that traits of mentioned language detected in it will be consistent elsewhere.

   Response: A lack of other comparable use-mention corpora does mean that the broad applicability of these results cannot yet be empirically tested. However, the conventional use of stylistic cues in Wikipedia to delimit mentioned language suggests that these findings will be applicable to other formally written texts as well. This use of stylistic cues is widely respected in writing, across several domains (Jr. Strunk and White 1979; Chicago Editorial Staff 2010; American Psychological Association. 2001). Below are some examples collected from other sources of text, with the original stylistic cues shown intact:

   - Like so many words, the meaning of “addiction” has varied wildly over time, but the trajectory might surprise you.[16]
   - Sending a signal in this way is called a **speech act**.[17]
   - M1 and M2 are Slashdot shorthand for “moderation” and “metamoderation,” respectively.[18]
   - He could explain foreordination thoroughly, and he used the terms “baptize” and “Athanasian.”[19]

[16] News article from CNN.com: `http://www.cnn.com/2011/LIVING/03/23/addicted.to.addiction/index.html`

[17] Page 684 of Russell and Norvig’s *Artificial Intelligence* (1995), a textbook.

[18] Frequently Asked Questions (FAQ) list on Slashdot, a news discussion website: `http://slashdot.org/faq/metamod.shtml`

- They use *Kabuki* precisely because they and everyone else have only a hazy idea of the word's true meaning, and they can use it purely on the level of insinuation.[20]

If the cues in grammar and vocabulary for mentioned language differed greatly between language sources (i.e., between Wikipedia and other texts), it would seem likely that the patterns of usage for stylistic cues would change radically as well. Their consistent application, in other words, seems to be a sign that the qualities of mentioned language are nominally consistent.

4.2.2 Corpus Creation

As stated in Section 4.2.1, one of the goals of this pilot study was to verify that stylistic cues are a practical mechanism for collecting instances of mentioned language. This section describes how the pilot corpus was created and then broken down into categories using the ostensive theory as a framework. Figure 4.1 outlines the creation of the pilot corpus, and the rest of this section describes it in detail.

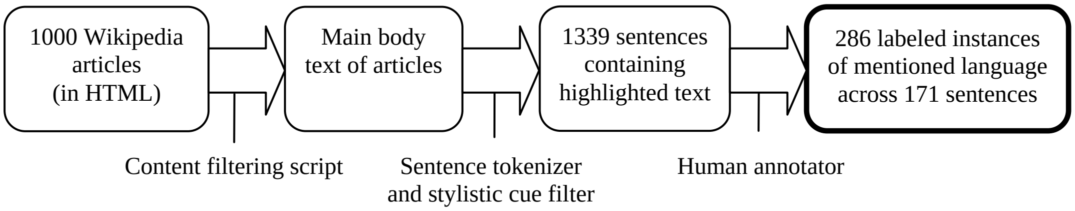
1000 Wikipedia articles (in HTML)
Main body text of articles
1339 sentences containing highlighted text
286 labeled instances of mentioned language across 171 sentences
Content filtering script
Sentence tokenizer and stylistic cue filter
Human annotator

Figure 4.1: The process of creating the pilot corpus.

[19] Novel *Elmer Gantry* by Sinclair Lewis.

[20] Opinion column on Slate.com: `http://www.slate.com/id/2250081/`

The annotation effort focused on 1000 randomly chosen articles from English Wikipedia. The articles were first processed in HTML format, since this was deemed the easiest medium for automated filtering. All information in article headers was discarded, as were end matter sections (such as “References”, “Notes”, “See Also”, etc.), since stylistic cues in them were often in non-sentential text. Tables and lists were also discarded for the same reason. Except for delimiters for bold and italic text, most of the markup left over from this procedure was removed, and the remaining text was segmented into sentences using NLTK’s (Bird 2006) implementation of the Punkt sentence tokenizer (Kiss and Jan Strunk 2011). These sentences were then filtered for those that contained *highlighted text*: that is, text in bold, italics, or between quotation marks. 1339 sentences contained one or more instances of highlighted text, and such instances became candidates for hand annotation.

Hand annotation required approximately three person-hours, with that time heavily skewed toward the first third of the sentences, as the set of categories was also developed during this labeling process. Only instances of highlighted text were considered for labeling, though multiple instances in each sentence were considered separately. Although only one researcher participated in the annotation effort, this was deemed acceptable since the labeling rubric from Chapter 2 was used and the pilot corpus would be superseded by later corpora with indications of robustness. Categories were formed around the appropriate *[item]* substitutions in the rubric (e.g., “this *proper name*”, “this *word*”, “this *symbol*”, “this *quotation*”), which are roughly analogous to ostentions of language, as introduced in Section 3.2.2.

4.2.3 Results and Discussion

Out of the 1339 sentences inspected by hand, 171 contained at least one instance of mentioned language. Many of those sentences contained several instances. Table 4.1 below lists the categories observed and their frequencies, and Table 4.2 shows examples from each category.

Category	**Code**	**Frequency**
Proper name	PN	119
Translation or Transliteration	TR	61
Attributed Language	AT	47
Words/Phrases as Themselves	WD	46
Symbols/Nonliteral Marks	SY	8
Phonetic/Sound	PH	2
Spelling	SP	2
Abbreviation	AB	1

Table 4.1: Frequencies of the categories of mentioned language found in the corpus. For brevity, the codes are used in Table 4.2 below.

Proper names were by far the most frequent category, with almost twice as many instances as the next most frequent category. This follows intuition, since Wikipedia articles often describe entities identified by proper names. In contrast, there were few instances of pronunciation (phonetic/sound) or spelling. Either the sentence filtering eliminated many instances of these before human annotation could find them, or they do not occur as frequently (at least in Wikipedia). Also noteworthy are the 46 instances of words or phrases as themselves, since these are examples of language being either introduced or clarified for the reader. While there exists a body of work on named entity recognition (Nadeau and Sekine 2007), very little exists on identifying when words serve a similar (but distinct) function as rigid designators for their types. One of the goals of this dissertation will be to fill that gap.

Category	Example
PN	In 2005, Ashley Page created another short piece on Scottish Ballet, a strikingly modern piece called The Pump Room, set to pulsating music by Aphex Twin.
TR	The Latin title translates as a method for finding curved lines enjoying properties of maximum or minimum, or solution of isoperimetric problems in the broadest accepted sense.
AT	It is still fresh in my memory that I read a chess book of Karpov by chance in 1985 which I liked very much, the 21-year-old said.
WD	Submerged forest is a term used to describe the remains of trees (especially tree stumps) which have been submerged by marine transgression, i.e. sea level rise.
SY	He also introduced the modern notation for the trigonometric functions, the letter e for the base of the natural logarithm (now also known as Euler's number) …
PH	The call of this species is a high pitched ke-ke-ke like American Kestrel.
SP	James Breckenridge Speed (middle name sometimes spelled Breckinridge) (1844-1912) was a successful businessman in Louisville, Kentucky and an important philanthropist.
AB	… Moskovskiy gosudarstvennyy universitet putej soobshcheniya, often abbreviated MIIT for Moscow Institute of Transport Engineers …

Table 4.2: Examples from the corpus of each category. Longer sentences for SY and AB have been truncated. Relevant instances of mentioned language appear as underlined, and the original stylistic cues have been removed.

4.2.4 Insights on Mentioned Language

The findings of this pilot study can be summarized in terms of the three goals listed in Section 4.2.1. Item numbers below correspond with the earlier numbering of the goals.

1) Stylistic cues are very good at delimiting the boundaries of mentioned language, but only nominally precise for retrieving instances of it. The researcher found very few instances where the sequence of mentioned words did not precisely correlate with the stylistic cues that highlighted them. Such accurate information on the boundaries of mentioned language will save time and effort for construction of later corpora. However, only 12.8% of the sentences that contained highlighted text actually contained mentioned language, and a higher percentage would speed up the necessary task of labeling the instances by hand.
2) Informally, several recurring "mention words" (e.g., *name*, *say*) were observed accompanying instances of mentioned language. These seemed likely to help in identifying instances of the phenomenon and expediting corpus creation. The next chapter will investigate this observation further.
3) The ostensive theory provided a useful framework for classifying instances of mentioned language, as it integrated easily with the labeling rubric to determine what non-use properties of language were being ostended. Still, the ostentions originally listed by Saka—however intuitive they seemed—were not evenly represented in the pilot corpus. PH, SP, and SY were some of the least common categories.

4.2.5 Bridge to the Combined Cues Study

This pilot study showed that it is possible to build a corpus of instances of mentioned language with a labeling rubric and with text sourced from Wikipedia. Stylistic cues were found to be a significant aid in the collection process, but with a precision of 0.128 the manual labeling task is still arduous and inefficient. Still, the pilot study revealed a new potential heuristic—metalinguistic cues in proximity to stylistic cues—and the next section will explore using this combination.

The next section presents the “combined cues” study, a second corpus-building effort that evolved from the lessons learned from the pilot study. In addition to stylistic cues, vocabulary cues were incorporated into sentence filtering prior to hand annotation. The section will describe these changes in procedure and the results of examining the composition of this second corpus.

4.3 The Combined Cues Study

The combined cues study inherits many aspects of the pilot study. Wikipedia was retained as a text source, since it performed as well as expected. The set of three stylistic cues (bold text, italic text, and text between quotation marks) were also retained, since they had a crucial role in collecting instances of mentioned language. Most details in article processing also remained the same.

However, in addition to the development of vocabulary heuristics (described in the next section), several other changes were deemed necessary. The combined cues study was intended to be *larger*, as prior efforts showed that it was likely that the procedure would produce meaningful results. Tables and lists in the article body were once more included, since their earlier exclusion had eliminated large amounts of potential text. However, sentences were required to contain at least 11 words to reach the stage of hand annotation, in order to eliminate unwanted non-sentential strings (e.g., short bulleted list entries, formatting fragments, various garbage-like fragments). Sentences that were entirely enveloped in a stylistic cue were also eliminated, as these lacked a proper context for labeling.

4.3.1 Candidate Collection and Labeling

The pilot study observed that instances of mentioned language are relatively sparse in Wikipedia article text, occurring on average less often than once per article. Since hand annotation was a necessary step in creating the ML corpus, some heuristics were used to gather a rich set of mentioned language candidates. Below, Figure 4.2 outlines the process of creating the corpus.

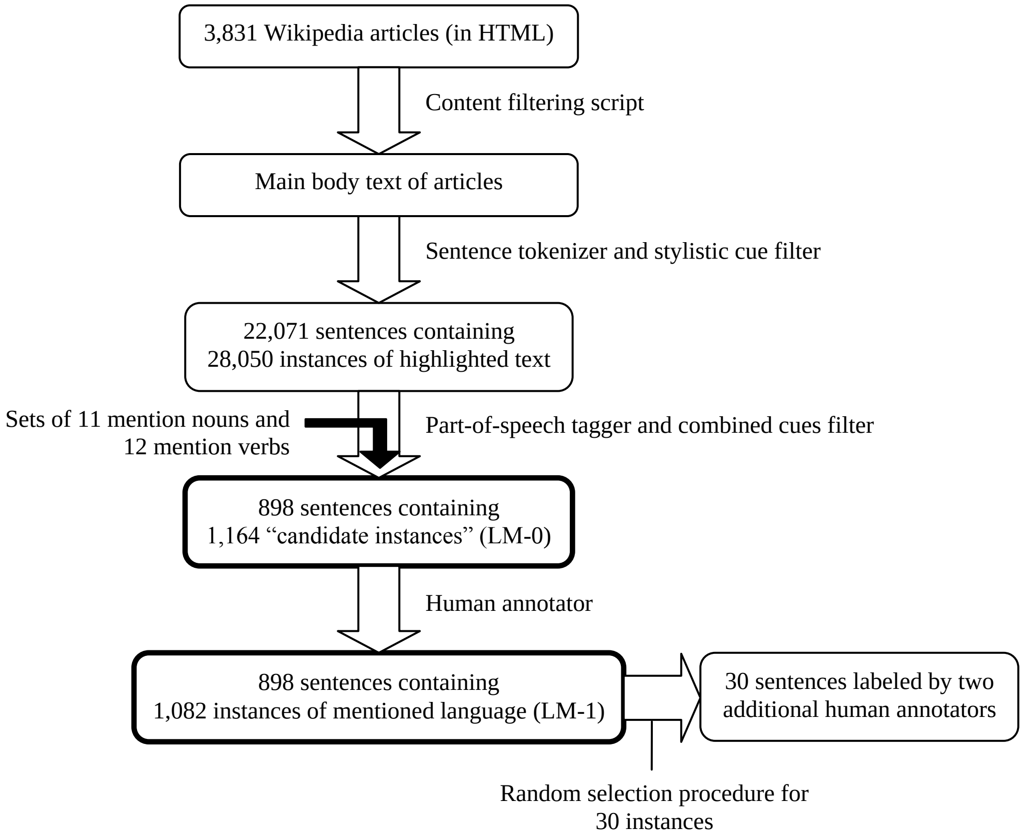
3,831 Wikipedia articles (in HTML)
Content filtering script
Main body text of articles
Sentence tokenizer and stylistic cue filter
22,071 sentences containing
28,050 instances of highlighted text
Sets of 11 mention nouns and
12 mention verbs
Part-of-speech tagger and combined cues filter
898 sentences containing
1,164 “candidate instances” (LM-0)
Human annotator
898 sentences containing
1,082 instances of mentioned language (LM-1)
30 sentences labeled by two
additional human annotators
Random selection procedure for
30 instances

Figure 4.2: The process of creating the Combined Cues Corpus.

Articles were randomly selected from English Wikipedia's most current article revisions, and heuristic filtering began at this level. Disambiguation pages were excluded from further examination, since they tend to be repetitive in structure and wording. Inside of articles, text from common end sections (i.e., "Sources", "References", "See also", and "External links") also was excluded, since text from those sources was frequently observed to be non-sentential. The remaining article text was then segmented into sentences using the Punkt sentence tokenizer (Kiss and Jan Strunk 2011). Those sentences that contained stylistic cues (bold text, italic text, or text between double quote marks) were retained, and all others were discarded. Applying this procedure to 3,831 articles produced a set of 22,071 sentences, which in turn contained 28,050 instances of text highlighted by stylistic cues (henceforth "highlighted text", for simplicity).

Initial examinations of these remaining sentences suggested that mentioned language occurred in fewer than one in ten of them, and an additional heuristic was applied beforehand annotation commenced. Using observations from the previous 171-sentence corpus, sets of "mention-significant" nouns and verbs were gathered. The appearance of a word from these sets near highlighted text signaled that the highlighted text was likely to be mentioned language. The procedure to gather these words was informal and manual, and a few potential mention-significant words (notably the verb *be*) were rejected because their great frequency reduced their significance as indicators. The eleven selected nouns and twelve selected verbs are listed below. The reader may note that most of the nouns refer to linguistic entities, while most of the verbs can serve as relational predicates or refer to speech acts:

Mention nouns: letter, meaning, name, phrase, pronunciation, sentence, sound, symbol, term, title, word

Mention verbs: ask, call, hear, mean, name, pronounce, refer, say, tell, title, translate, write

Words in the sentences were part-of-speech tagged and stemmed, again using tools from NLTK. The sentences were then filtered for those in which a mention word occurred (respecting the part of speech of its set) in the three-word phrase preceding text highlighted by a stylistic cue. This resulted in a set of 898 sentences, which in turn contained 1,164 instances of highlighted text. This set of instances was named the *ML-0 set*.

Manual annotation of mentioned language then commenced. To eliminate possible biases, all three stylistic cues were substituted with pairs of asterisks (delimiting the beginning and ending of highlighted text) prior to inspection. The researcher then considered each instance in the ML-0 set and decided if it was mentioned language by reading its containing sentence and applying the rubric from Chapter 2. 1,082 instances were deemed to be mentioned language, and this set was named the *ML-1 set*, which also serves as the *ML corpus*. This figure suggests that the heuristics leading to the creation of the ML-0 set have approximately 93% precision for retrieving mentioned language, though their recall has not yet been measured.

4.3.2 Reliability and Consistency

Another limitation of the ML corpus is the lack of participation from multiple readers. To explore the possible impact of this, two additional human readers worked separately (from each other and from the primary reader) to annotate a 30-instance subset of the ML-0 set. These readers were also well-acquainted with the detection of mentioned language. Half of the 30 instances were selected from those annotated by the primary reader as mentioned language, and half were selected from those annotated as not. With that condition, the instances were randomly chosen from the ML-0 set, shuffled, and then distributed to the additional readers.

All three readers produced the same annotation for 25 of the instances, and on each of the remaining five instances, the additional readers differed with each other. (Since the annotation scheme was binary, this meant that one additional reader agreed with the primary reader and one disagreed). The kappa statistic (Cohen 1960) was 0.779. These results were taken as a mild indication of the reliability and consistency of annotations in the ML corpus.

4.3.3 Corpus Composition

This section will present some notable findings distilled from the ML-0 and ML-1 sets. Particular attention was given to the precision of the heuristics used to create the ML-0 set. The combination of heuristics performed better (at 93% precision overall) than had been expected, with some standout performances from specific mention words and stylistic cues.

Below, Table 4.3 shows the frequency of mention words in the three-word phrases preceding each instance (an instance being a string of highlighted text) in the ML-0 set. Mention words were only counted if they appeared as their set-appropriate parts of speech. In the tables in this section, the precision shown is the percentage of those instances deemed by the primary human reader to be mentioned language and thus placed in the ML-1 set.

Mention word	**Frequency**	**Precision (%)**
call (v)	349	98.6
name (n)	153	98
name (v)	89	94.4
say (v)	86	94.2
term (n)	79	98.7
title (n)	72	84.7
title (v)	64	96.9
word (n)	55	100
write (v)	52	50
mean (v)	39	100
refer (v)	35	85.7
meaning (n)	20	100
translate (v)	20	20
phrase (n)	18	100
symbol (n)	10	80
pronounce (v)	8	100
tell (v)	7	71.4
letter (n)	6	33.3
pronunciation (n)	4	100
ask (v)	4	75
sentence (n)	3	33.3
hear (v)	3	0
sound (n)	1	0

Table 4.3: Frequencies of mention nouns (n) and verbs (v) in the three words preceding each instance in the ML-0 set, with their precisions for retrieving mentioned language.

The verb *call* and the noun *name* stood out as the most common of the mention words, with all others forming a relatively smooth tail of descending frequency. Both words also had substantially above-average precision. *Word* (n), *meaning* (n), *phrase* (n), *pronounce* (n), and *pronunciation* (v) all had perfect precision, though they appeared less frequently. However, following the multiple-reader experiment in Section 4.2.2, it was discovered that meaning instances were particularly difficult to classify, generating some debate among the participants. Finally, an observant reader may note that the frequencies in Table 4.3 sum to 1,177 instead of 1,164 (the size of the ML-0 set). This is because 13 instances had more than one mention word in the preceding three-word phrase. All 13 of these instances were annotated as mentioned language.

Although stylistic cues were hidden from the readers while they annotated instances, data on the cues was retained. Table 4.4 breaks down their frequencies and precisions. Double quote marks had the highest frequency, and the reason was first assumed to be frequent quotation (in the sense of speech reporting, for example) in Wikipedia. However, as the next table will show, that was probably not the case. Italics had by far the lowest precision. 23 of the 58 non-mention italic instances had *write* (v) as a preceding mention word, which conjures a common construction (as in "Dickens wrote *Great Expectations*…") that does not involve mentioned language. Bold had both the highest precision and lowest frequency. It is worth noting that Wikipedia articles, by convention, contain the article subject in bold text in the first sentence.

Stylistic cue	**Frequency**	**Precision (%)**
double quote	601	96.7
italic	427	86.4
bold	136	97.1

Table 4.4: Frequencies of stylistic cues in the ML-0 set and their precisions for retrieving mentioned language.

Prior to analysis, it was hypothesized that the proximity of a mention word to highlighted text increases its likelihood of being mentioned language. Table 4.5 shows this hypothesis to be true, albeit in the limited three-word window that was examined. Also shown are overall frequencies and precision percentages (weighted by frequencies) for nouns and verbs. There appears to be a strong correlation between proximity and precision, though proximity in this data does not account for the grammatical structure of corpus sentences, which will deserve examination in further research. A mention verb directly preceding highlighted text was by far the most common combination. Overall, mention nouns had a slightly greater precision than mention verbs.

Noun/Verb position	**Frequency**		**Precision (%)**	
	Noun	**Verb**	**Noun**	**Verb**
1	281	458	98.6	97.2
2	89	179	91.0	85.5
3	51	119	76.5	84.0
overall	421	756	94.3	92.4

Table 4.5: Frequencies of mention nouns and verbs in the three words preceding highlighted text (e.g., word position 1 is the word just before the highlighted text), with their precisions for retrieving mentioned language.

Finally, Table 4.6 shows the most common mention word-stylistic cue combinations in the ML-1 set. The prevalence of *call* (v) is once again apparent, as its combinations with double quote marks and italics have a substantial lead in frequency over all other combinations. Double quote marks with *say* is the third most common combination, which matches earlier intuitions on quotation, but the same stylistic cue appears frequently with *call* (v), *name* (n), *term* (n), and *name* (v) as well. Bold makes only one appearance in the top ten, in combination with the previously mentioned *call* (v). These ten combinations account for only 17% of the distinct combinations observed but 62.6% of all instances in the ML-1 set.

Word	**Cue**	**Frequency**	**% of total**
call (v)	d. quote	151	14.0
call (v)	italic	133	12.1
say (v)	d. quote	74	6.8
name (n)	italic	60	5.5
name (n)	d. quote	56	5.2
call (v)	bold	53	4.9
term (n)	d. quote	45	4.2
name (v)	d. quote	39	3.6
title (v)	italic	36	3.3
title (n)	italic	32	3.0

Table 4.6: The ten most frequent word and stylistic cue combinations in the ML-1 set, with their percentages of the total (1082) instances. Out of 69 possible different word-cue combinations, 59 were observed.

4.3.4 Observations

Overall, these results seem to validate the heuristics that were used to collect candidate instances. They also further confirm that Wikipedia is a fertile source of mentioned language, as the instances in the ML-1 set exhibit a variety of different constructions. Given the size of Wikipedia and the current methods for collecting candidates, running a final, more refined iteration of this procedure will be both possible and desirable.

The combined stylistic and vocabulary cues had a precision of 93% in retrieving mentioned language, which was higher than the researcher had expected and perhaps an indication that the vocabulary filtering was *too* selective. Although the list of 13 selected mention words contains many common metalinguistic words, it is unlikely to be an exhaustive list of them. Further work will be necessary to gather more mention words, perhaps by using the existing list to seed a search in a lexicon for related ones. The next chapter will explore this idea.

Finally, the above results hint at some patterns in how stylistic cues occur in text. Although the correlation between quotation and quotation marks is well-known and perhaps obvious in their names, the exact reasons why writers use bold and italics are not as clear. Opportunities exist here for research that would benefit copyediting and publication technologies, though they fall beyond the scope of this dissertation.

This section presented an intermediate step in the creation of a robust corpus of mentioned language. The combination of lexical and stylistic cues was effective at retrieving a rich set of instances of mentioned language, though possibly at the expense of lower recall, since the sought vocabulary set was relatively small. The next (and final) corpus creation study will attempt to strike a balance between time-consuming manual annotation and all-too-effective automated filtering methods.

4.4 The Enhanced Cues Study

4.4.1 Motivation

The Combined Cues Study showed that the combination of lexical cues and stylistic cues can be used to render practical the task of creating a corpus of mentioned language. The inter-annotator agreement experiment which accompanied it demonstrated that, while perfect agreement on labeling all candidate instances is unlikely, high agreement does exist among skilled annotators. With over 1,000 labeled instances, this second study was much larger than the Pilot Study, and some additional patterns in mentioned language were discernable in the data.

However, the Combined Cues Study was intended as an intermediate step in the creation of a more robust corpus, and it had several limitations. The vocabulary set used to filter instances of highlighted text consisted of just 23 words, and although these were gathered empirically from the pilot study, it seems likely that far more mention words exist. The filtering procedure only examined words before highlighted text, and substantial lexical cues sometimes occur *afterward* as well (as in sentences like "*Marathon* refers to a town in the Florida Keys", where the verb *refer* is a significant cue). The filtering of instances was far more precise than expected, even to a fault, as it introduced concerns about recall. Given the simple nature of the filtering and the limited vocabulary set, it seemed unlikely that a procedure that produced a 93%-positive mixture of instances was able to capture a robust sampling of mentioned language in the original Wikipedia article text. This unbalanced mixture also posed a problem for future machine learning efforts: without a larger set of verified negative instances, it would be difficult to train and evaluate a classifier.

These limitations motivated the construction of a third and final corpus of mentioned language. Wikipedia was retained as a source of text, since it had performed well in the previous two studies. The third corpus inherited many characteristics from the intermediate study, such as the basic method of filtering sentences and the structure of the inter-annotator agreement study on a subset of the data. It also borrowed some inspiration from the pilot study, in the form of phenomenon categories and a broad focus on the varieties of mentioned language. The next section will explain the major differences between this third study and those previous efforts.

4.4.2 Changes from the Previous Studies

Experiences and results from the Pilot Study and the Combined Cues Study influenced the structure of this third study. Overall, the goal was no longer to verify methods for collecting instances of mentioned language; instead, it was to create a corpus of instances that represented the breadth of mentioned language in Wikipedia and would be conducive to efforts to detect mentioned language automatically. One particularly important change was the use of *WordNet* to gather a large, varied set of words to serve as lexical cues. This section describes this use of WordNet and details other important departures from the previous studies.

The Combined Cues Study used a set of eleven nouns and twelve verbs to help filter instances of highlighted text (i.e., words in bold text, italic text, or text between quotation marks). When one of these "mention words" was found in the three words preceding highlighted text, the containing sentence was deemed a "candidate" and examined by a human reader. Although this set of words was chosen with careful thought, it was not intended to be *complete* so much as *effective*. A more comprehensive set was desirable for the third study, since it would lead to a wider coverage, and a resource was sought to expand the collection of mention words.

WordNet (Fellbaum 1998), a lexical database for the English language, was selected to help gather additional mention words. WordNet organizes *lemmas* (words in their canonical or "dictionary" forms) into sets of synonyms called *synsets*, and these synsets are linked to each other through semantic relationships. These relationships include hypernymy, hyponymym, holonymy, and troponymy, though the terms for the relationships and the types of relationships available vary depending on the part-of-speech of the synsets. Nouns, verbs, adjectives, and adverbs are represented in the database in separate, connected graphs. Because of this structure, WordNet can be thought of as an ontology of words.

The mention words for the second study were used as a "seed" to begin a search for synsets in WordNet that contained lemmas likely to co-occur with mentioned language. The search procedure was divided into two stages, the first based upon human effort and the second a "brute force" automated crawl. It was designed to be as inclusive as possible, even at the expense of gathering some lemmas that were unlikely to have metalinguistic value. Figure 4.3 below illustrates the procedure for the noun *term*.

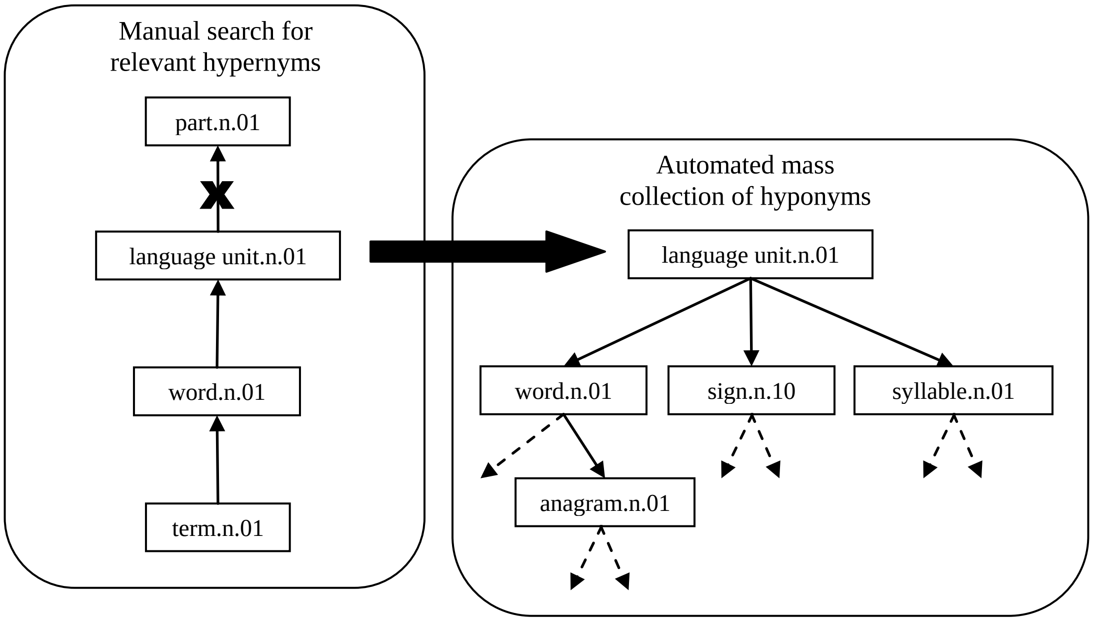

Figure 4.3: Illustration of the two stages of the WordNet search for mention words, seeded by the noun *term*. Each synset is labeled with a lemma ("term"), a part of speech ("n") and a sense number ("01") Note that *X* is a hypernym of *Y* if every *Y* is a kind of *X*; hyponymy is the converse of this relationship.

In the diagrammed example, the first stage began with the researcher reading the definitions of synsets containing the lemma *term*. Those synsets with definitions that indicated firm metalinguistic use were retained, while the rest were discarded. Then, all hypernym link(s) (direct and inherited) from each relevant synset were followed until synsets were reached that were not directly related to metalanguage. In Figure 4.3, this break occurs between *language unit* and *part*. Holonym ("*Y* is part of *X*") and meronym (the converse) relationships were also examined and followed. Applying this procedure to all 23 of the original mention words produced a set of 111 relevant synsets. The increase in count was due largely to each mention word occurring in several synsets (indicating several senses for a word), though sometimes these starting points "merged" through mutually inherited hypernyms.

In the second stage, all hyponym links (direct and inherited) from the 111 relevant synsets were followed in an exhaustive, automated crawl of the graph, and all lemmas discovered during the crawl were gathered into a list. Both single words and co-locations (e.g., "chew out") were collected. With duplicates removed, this method produced a list of 8,735 unique strings, comprising about 6.8% of all unique strings in the noun and verb synsets of WordNet. Among the list were many strings that seemed unlikely to co-occur significantly with mentioned language, but they were retained, since selectively removing them would have been both time-consuming and difficult. This list served as the third study's equivalent to the list of mention words used in the Combined Cues Study.

In addition to the use of WordNet, the third study differed from the second study in several other ways, listed below.

- In the second study, the three words prior to highlighted text were scanned for mention words; in the third study, both the three words before *and* the three words after highlighted text were scanned.
- Prior to human annotation, sentences in the third study were filtered heuristically to reduce the number of instances of highlighted text that were likely to be speech acts and proper titles. This was done because those two categories dominated the Pilot Study, and for practical applications it was desirable to increase the number of instances of other categories, particularly those of a “pure mention” nature (e.g., introducing terms and symbols).
- Other heuristics were used to limit the collection of sentences that lacked sufficient context for human readers to confidently label. Highlighted text inside of parentheses often suffered this problem in the second study, as did italicized mathematical symbols; both were excluded from becoming candidate instances in this third study.
- When mention words occurred inside of highlighted text, they were not considered to be cues (and thus could not be used to promote a second nearby instance of highlighted text to candidacy). This change would have had little impact in the second study, but it was necessary in the third study since the list of mention words became much larger.
- Categories for mentioned language, which were used in the first corpus, were reintroduced, though they were modified to better correspond with how the phenomenon was observed to occur. These categories will be described in the next section.

4.4.3 Corpus Creation

At a high level, the process of creating the Enhanced Cues Corpus was similar to the procedure used to create the Combined Cues Corpus, with the particular changes listed previously. This section will describe the process, which is outlined in Figure 4.4 below.

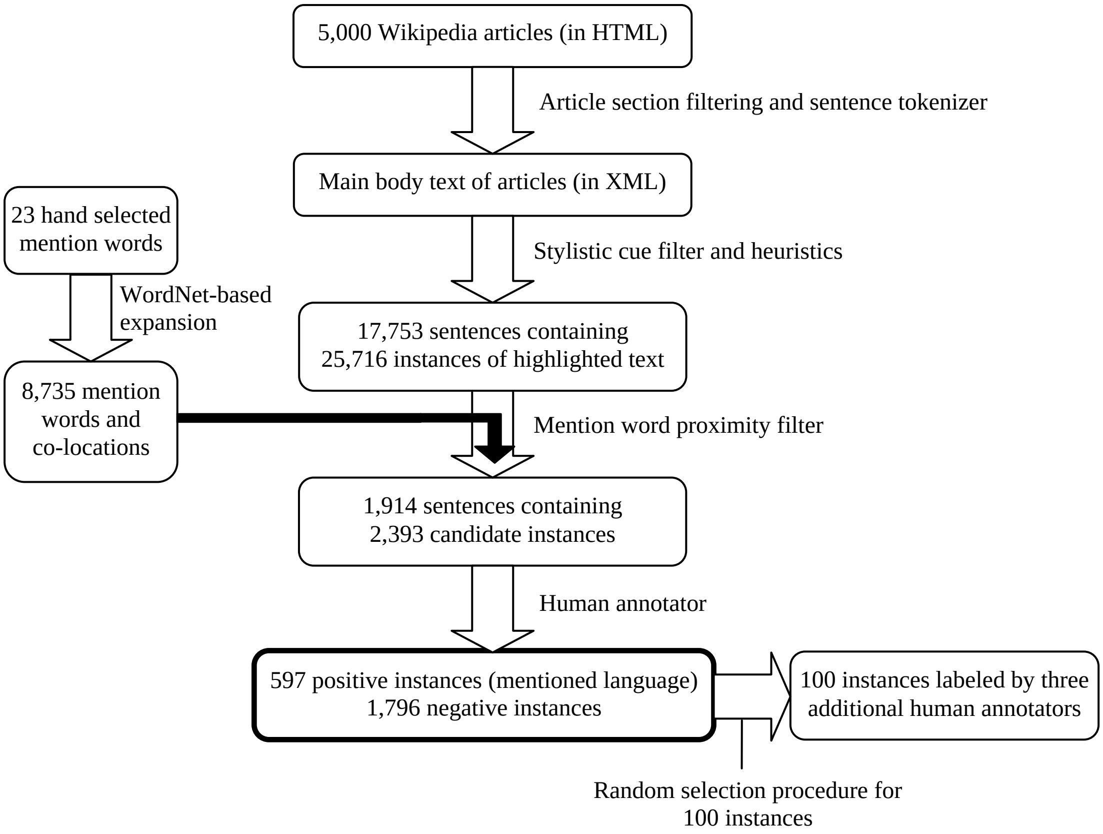
5,000 Wikipedia articles (in HTML)
Article section filtering and sentence tokenizer
Main body text of articles (in XML)
Stylistic cue filter and heuristics
23 hand selected mention words
WordNet-based expansion
17,753 sentences containing 25,716 instances of highlighted text
8,735 mention words and co-locations
Mention word proximity filter
1,914 sentences containing 2,393 candidate instances
Human annotator
597 positive instances (mentioned language) 1,796 negative instances
100 instances labeled by three additional human annotators
Random selection procedure for 100 instances

Figure 4.4: The procedure used to create the Enhanced Cues Corpus.

First, 5,000 articles were randomly selected from English Wikipedia's most current article revisions, excluding disambiguation pages. Article contents were saved in HTML first and then reprocessed into a XML format designed to ease bookkeeping of data and to decrease the work necessary for future related research efforts. The

XML format segments and enumerates the sentences in the main body of each article (discarding sections such as “Sources”, “References, “See Also”, and “External Links”), and also enumerates each instance of highlighted text inside of each sentence. As in the previous two studies, *highlighted text* is considered to be text in bold, in italics, or between double quotation marks.

To gather a rich set of candidate instances, sentences containing highlighted text were subject to several filters. The heuristics for pure mention and sufficient context, described in the previous section, were applied at this stage. Next, as in the second study, candidate instances were identified. An instance of highlighted text was promoted to a candidate instance if it had one of the 8,735 mention words (or co-locations) nearby; for this study, mention words were searched for in the three words before *and* after highlighted text. Out of the 25,716 total available instances of highlighted text in 17,753 sentences, 2,393 candidate instances were identified across 1,914 sentences.

The candidate instances were then annotated by the researcher, requiring approximately four hours. As in the second study, delimiters of stylistic cues were replaced by asterisks to eliminate possible biases. The human reader considered each candidate instance within the context of its sentence and labeled it using the rubric in Chapter 2 and the practical guidelines in Appendix A. These guidelines were intended to encourage uniformity between the researcher’s annotations and those of the additional annotators, and they reflected observations made during the previous two studies. Each candidate instance received one of five labels, with descriptions here reproduced from the guidelines:

1) *Words as Words (WW)*: Within the context of the sentence, the starred phrase is used to mean the word or phrase itself and not what it usually refers to.
2) *Names as Names (NN)*: The sentence directly refers to the starred phrase as a proper name, nickname, or title.
3) *Spelling or Pronunciation (SP)*: The starred text appears only to illustrate spelling, pronunciation, or a character symbol.
4) *Other Mention/Interesting (OM)*: Instances of mentioned language that do not fit the above three categories.
5) *Not Mention (XX)*: The starred phrase is not mentioned language.

Examples for each of these (as well as related negative examples) can be found in Appendix A. Only candidate instances were given labels, though occasionally "non-candidate" instances of mentioned language were observed in sentences. 597 candidate instances were labeled WW, NN, SP, or OM, and the remaining 1,796 were labeled XX, producing a ratio of about 1:3 between positive and negative instances.[21]

4.4.4 Annotation Results

As previously stated, about 25% of the candidate instances—those segments of highlighted text with at least one mention word in the three words before or after—were deemed to be mentioned language. This section will examine the composition of the corpus, with special attention to the occurrences of those mention words. Table 4.7 breaks down the corpus by the labeling categories, and Table 4.8 presents two examples of each category from the corpus. WW was by far the most common category of mentioned language to appear, indicating that the heuristics designed to

[21] Note that these figures are not directly comparable to either of the previous two corpora due to the changes in heuristics, mention words, and pre-candidate filtering.

favor instances of “pure” mention had the desired effect. As in the Pilot Study, instances of spelling and pronunciation were particularly hard to find. The OM category was occupied mostly by instances of language production by agents, such as those shown for the category in Table 4.8.

Category	Code	Frequency
Words as Words	WW	438
Names as Names	NN	117
Spelling or Pronunciation	SP	48
Other Mention/Interesting	OM	26
Not Mention	XX	1764

Table 4.7: The composition of the candidate instances by category as labeled by the researcher.

In the Combined Cues Study, mention words were counted in text only if they appeared in their preselected parts of speech, but in this third study no such restriction was enforced. Still, in the interest of gathering both lexical and grammatical cues for mentioned language, part-of-speech tags were computed for words in all the candidate instances. Tables 4.9 and 4.10 below list the ten most common words (as part-of-speech-tagged) before and after (respectively) candidate instances. Many of them were in the list of mention words for this third study (and the list for the second study, even), but some were new.

Category	Example
WW	The IP Multimedia Subsystem architecture uses the term transport plane to describe a function roughly equivalent to the routing control plane. The material was a heavy canvas known as duck, and the brothers began making work pants and shirts out of the strong material.
NN	Digeri is the name of a Thracian tribe mentioned by Pliny the Elder, in The Natural History. Hazrat Syed Jalaluddin Bukhari's descendants are also called Naqvi al-Bukhari.
SP	The French changed the spelling to bataillon, whereupon it directly entered into German. Welles insisted on pronouncing the word apostles with a hard t.
OM	He kneels over Fil, and seeing that his eyes are open whispers: brother. During Christmas 1941, she typed The end on the last page of Laura.
XX	NCR was the first U.S. publication to write about the clergy sex abuse scandal. Many Croats reacted by expelling all words in the Croatian language that had, in their minds, even distant Serbian origin.

Table 4.8: Two examples from the corpus for each of the categories. The candidate instances appear underlined, with the original stylistic cues removed.

Finally, while the Combined Cues Study treated mentioned language as a largely homogenous phenomenon, this third study reintroduced categories, and differences between the vocabularies associated with each category were apparent. Table 4.11 breaks down instances in each category in the same manner as Tables 4.9 and 4.10. Note that XX (Not Mention) is included for completeness. Due to sample size, several combinations of category and position had little recurring vocabulary. Others, such as the sought-after WW, were richer. This and other issues will be discussed further in Section 4.4.6.

Rank	Word	Frequency	Precision (%)
1	call (v)	92	80
2	word (n)	68	95.8
3	term (n)	60	95.2
4	name (n)	31	67.4
5	use (v)	17	70.8
6	know (v)	15	88.2
7	also (rb)	13	59.1
8	name (v)	11	100
9	sometimes (rb)	9	81.9
10	Latin (n)	9	69.2

Table 4.9: The top ten words appearing in the three-word sequences *before* candidate instances (combined with their simplified part-of-speech tags), with their frequencies and precisions in retrieving mentioned language.

Rank	Word	Frequency	Precision (%)
1	mean (v)	31	83.4
2	name (n)	24	63.2
3	use (v)	11	55
4	meaning (n)	8	57.1
5	derive (v)	8	80
6	refers (n)	7	87.5
7	describe (v)	6	60
8	refer (v)	6	54.5
9	word (n)	6	50
10	may (md)	5	62.5

Table 4.10: The top ten words appearing in the three-word sequences *after* candidate instances (combined with their simplified part-of-speech tags), with their frequencies and precisions in retrieving mentioned language.

Category	Rank	Before Instance		After Instance	
		Word	Frequency	Word	Frequency
WW (438)	1	call (v)	70	mean (v)	28
	2	word (n)	67	use (v)	9
	3	term (n)	60	meaning (n)	7
	4	use (v)	14	refers (n)	7
	5	name (n)	12	describe (v)	6
NN (117)	1	call (v)	24	name (n)	18
	2	name (n)	18	mean (v)	3
	3	name (v)	7	spell (v)	2
	4	title (n)	5	also (rb)	2
	5	nickname (v)	4	derive (v)	2
SP (48)	1	letter (n)	6	use (v)	2
	2	write (v)	3	result (n)	2
	3	spelling (n)	3	two (cd)	2
	4	contain (v)	2	codename (n)	1
	5	spell (v)	2	smallish (v)	1
OM (26)	1	call (v)	10	say (v)	2
	2	say (v)	2	add (v)	1
	3	ethic (n)	1	call (v)	1
	4	refers (n)	1	despite (in)	1
	5	type (v)	1	decide (v)	1
XX (1764)	1	mean (v)	45	commune (n)	38
	2	describe (v)	24	also (rb)	18
	3	consider (v)	19	name (n)	14
	4	name (n)	15	write (v)	13
	5	say (v)	14	call (v)	13

Table 4.11: By category, the top five words appearing in the three word sequences respectively *before* and *after* candidate instances (combined with their simplified part-of-speech tags). Numbers in parentheses below the category labels indicate their frequencies (repeated from in Table 4.7).

4.4.5 Inter-Annotator Agreement Study

To provide some indication of the reliability and consistency of the Enhanced Cues Corpus, three additional human readers were recruited to label subsets of the candidate instances. Participants were given the guidelines in Appendix A and first assigned a “trial run” task to label the same 10 sentences. These were hand-picked to reveal the effectiveness of the annotation guidelines, which were adjusted slightly afterward for clarity. Then, participants were given a set of 100 sentences to label (again, the same sentences for all participants, though shuffled differently for each participant). These consisted of instances randomly selected to fill quotas from each category. For both tasks, annotators worked independently from each other and from the researcher. Table 4.12 shows the by-category breakdown of instances selected for these two annotation efforts.

Category	**Code**	**Trial Run**	**Final Run**
Words as Words	WW	2	17
Names as Names	NN	2	17
Spelling or Pronunciation	SP	2	16
Other Mention/Interesting	OM	0	4
Not Mention	XX	4	46

Table 4.12: The composition of the sets of sentences given to annotation participants for the Trial Run and the Final Run. In the Trial Run, sentences were hand-picked inside each category, and in the Final Run, sentences were randomly chosen inside each category.

The by-category compositions (i.e., how many were chosen from each category) for the Trial Run and the Final Run were not chosen to be representative of the population of candidate instances. Rather, they were chosen to illustrate whether the researcher’s labels for each category would correspond with participants’ labels. Spelling and Pronunciation, for instance, was scarcely found among candidate instances, and “scaling down” the researcher’s category counts from the candidate instances (presented in the next section) would have resulted in very few sentences from the SP category for participants to label.

The three additional annotators were asked to approximately keep track of the time they needed to label the 100-sentence set, and their self-reported times were 20 minutes, 30 minutes, and 30-45 minutes. Based on these durations, a plan to have them label the rest of the candidate instances was deemed infeasible. Still, the data gathered from the 100-sentence set provided some insight into the repeatability and consistency of mentioned language labeling.

A series of calculations of the kappa statistic (Cohen 1960) revealed the structure of label concurrence between the participants. First, calculations were done to determine the level of agreement on the mere presence of mentioned language, by mapping labels *WW*, *NN*, *SP*, and *OM* to *true* and *XX* to *false*. Under this scheme, K among all four annotators (the primary annotator and the three additional) was 0.74. All four annotators agreed upon a *true* label for 46 instances and a *false* label for 30 instances. Under the same relabeling scheme, K between the primary annotator and a hypothetical “majority voter” of the three additional annotators was 0.90. These results were taken as a mild to moderate indication of the reliability of the “simple” mention vs. non-mention labeling of the full set of candidate instances.

However, the per-category results showed reduced levels of agreement. K for the original category labels was 0.61, and Table 4.13 lists K for labeling schemes "binarized" with respect to each category.

Category	WW	NN	SP	OM	XX
K	0.38	0.72	0.66	0.09	0.74

Table 4.13: Values of the kappa statistic for category-based binary relabelings of candidate instances. For each category label, all other labels were mapped to a single *other* label, and kappa was then calculated on the remapped annotations.

A low K-value for remapped OM was expected, as the instructions for the category were vague. The primary annotator and two of the three additional annotators used it with roughly the same frequency but generally on different instances; the remaining annotator declined to use it at all. The binary relabeling with respect to XX is equivalent to the *true-false* remapping previously discussed. K values for remapped WW, NN, and SP labels were substantially lower than K for the original labels or for the *true-false* remapping. This contrast suggests that, although annotators tend to agree whether a candidate instance is mentioned language, there is less of a consensus on how to further qualify positive instances.

4.4.6 Discussion

The list of the most common words to appear before candidate instances bears a resemblance to the list of mention words from the Combined Cues Study, as the top four words in Table 4.13 are in the top eight most frequent mention words from the second study (see Table 4.3). Even though the list of mention words for this third study was much longer and heuristics were used to bias the candidate selection, these four words still frequently occurred with mentioned language, suggesting that they are core metalinguistic terms. *Use* (v) and *know* (v) also appeared often before mentioned language, and it seems intuitive that they too are common metalanguage. Puzzlingly, 81.9% of the appearances of *sometimes* before candidates occurred with mentioned language; this might be an artifact of common tendencies in Wikipedia writing (e.g., phrases like "sometimes called…" occurring often). The most common words after candidate instances (in Table 4.10) also featured many terms from the second study's list of mention words.

Although the recurrence of mention words from the Combined Cues Study was a promising development, the overall decrease in their precisions for retrieving mentioned language was unexpected. With the exception of *name* (v), all of the precisions of the recurring words fell, with the greatest reduction being *name* (n) at 30%. No certain cause for this is known, though the researcher hypothesizes that it is a result of the intentional heuristic bias against instances of mentioned titles and proper names. It is possible that instances in the NN category have much stronger lexical cues than the other categories of mentioned language, and in their relative absence, overall precisions fell. Notably, many of them still remained above 80%.

The distributions of word frequencies both before and after candidate instances showed "long tail" behavior beyond the first few items. Tables 4.9 and 4.10 begin to demonstrate this, and it was further evident in the full data used to construct them but not shown here. The by-category breakdown of vocabulary in Table 4.11 shows the same trend, though this might be partially due to sparseness in the SP and

OM categories. The analysis of the Combined Cues Study did not demonstrate this quality, since the focus was only on the frequencies of selected words. Thinking toward the automatic detection of mentioned language, it seems promising that most of the highest frequency words appear to be metalanguage. Still, it seems likely that there is a much larger set of metalinguistic cues that occur only sparsely in text, and such sparseness could limit the recall of mentioned language detection.

4.4.7 Conclusions

Section 4.4 described the third and final corpus of mentioned language created for this dissertation project. It incorporates properties of both of the previous two corpora to enhance its coverage, robustness, and suitability for future research. The vocabulary sieve for candidate instances was greatly widened, allowing for a much greater variety of constructions of mentioned language to be gathered. The balance of the corpus composition was tipped in favor of instances of "pure mention", since they are of particular interest both practically and historically in the study of the use-mention distinction. The corpus will enable the remaining goal of this dissertation, the automatic detection of mentioned language in text, discussed in the next chapter.

Chapter 5: Detection and Delineation of Mentioned Language

5.1 Introduction

This chapter presents some results on the feasibility of detecting and delineating mentioned language in text without the aid of stylistic cues, which have served as a "crutch" for building the corpora in previous chapters. Metalinguistic cues, syntactic patterns, and semantic roles are shown to be significant aids for the problem. Although it was not possible to comprehensively detect and delineate all mentioned language, the methods discussed in this chapter show promise for broad coverage of the phenomenon.

5.2 Motivation

One of the goals of this dissertation project is to develop methods of detecting and delimiting mentioned language, and the corpora built up to this point have begun to demonstrate how that goal might be accomplished. The co-occurrence of "mention words" with stylistic cues has been a useful pattern for both detecting and delineating the phenomenon, but stylistic cues are unlikely to be dependable or present at all in many sources of text or other media. While bold text, italic text, and quotation marks seem to be natural markers for mentioned language, they are subject to inconsistent standards of usage, uneven application of standards, and difficulties integrating the cues into language analysis. Most of the practical applications of detecting mentioned language (listed in Section 1.2) cannot rely upon stylistic cues. Although the cues were a necessary "crutch" for studying mentioned language, the detection of the phenomenon must move beyond them.

Without stylistic cues, the problem of detecting mentioned language becomes more complicated. The “candidate instances” of the corpus studies no longer exist; now, any sentence might contain mentioned language, and any substring of words within a sentence might be an instance of mentioned language. These two identification problems correspond respectively with the *detection* and *delineation* tasks described in Section 1.3, and both must be accomplished to fully predict an occurrence of mentioned language. Observations made while building the corpora in previous chapters suggest that such predictions will be possible, though variations in the natural occurrences of the phenomenon may limit performance. Common constructions of mentioned language are likely to be easily detectable, while some portion of the “long tail” of less common constructions may require semantic analysis beyond the capabilities of existing language technologies.

Beyond the analysis in Section 4.4, further examination of the Combined Cues Corpus showed some trends in vocabulary that served as a starting point for both the detection and delineation tasks. Figures 5.1 and 5.2 below illustrate them. Together, the four words that most frequently preceded candidate instances—*call* (v), *word* (n), *term* (n), and *name* (n)—covered 42% of those instances that were mentioned language and just 2.4% of those that were not. A similar trend of metalinguistic terms providing greater coverage for mentioned language was observed for words following candidate instances, though it was not as dramatic; mean (v), name (n), use (v), and meaning (n) covered only 12% of positive instances and 1.9% of negative instances. Both figures illustrate a “long, thin tail” behavior of metalanguage: a few very frequent words cover many instances of mentioned language, while the many remaining words cover just a few instances each. If methods exist to detect and delineate mentioned language based on the most common metalinguistic cues, those methods would cover a substantial subset of positive instances.

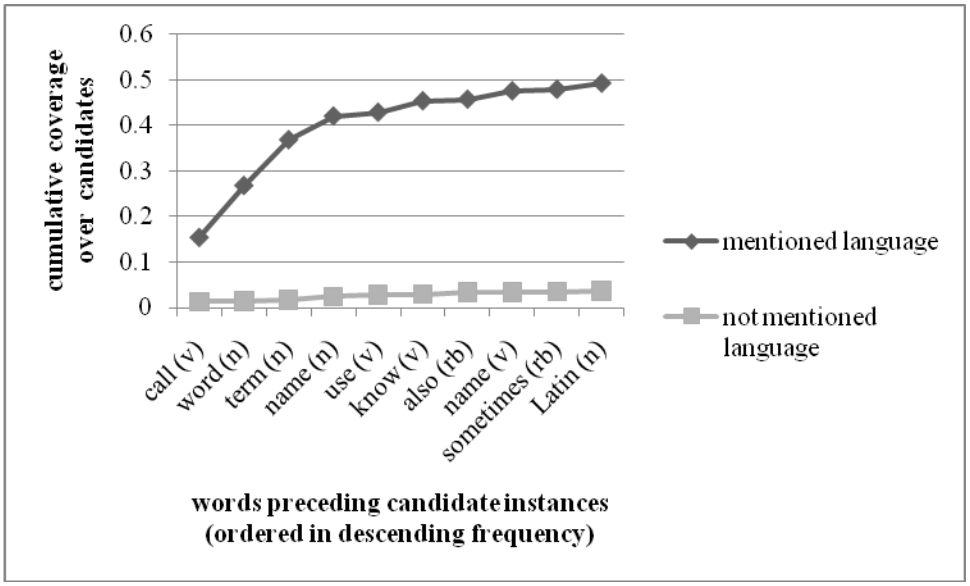

Figure 5.1: Cumulative coverage over candidate instances by the most common words to precede candidate instances in the Enhanced Cues Corpus. For example, *call* (v) or *word* (n) appeared in the three words preceding 27% of candidate instances that were mentioned language and 1.4% of those that were not mentioned language.

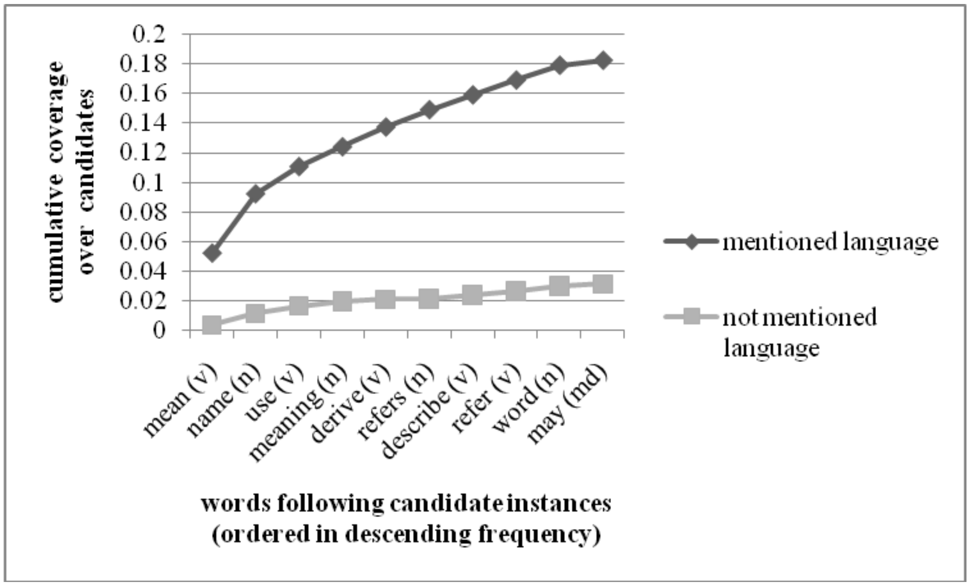

Figure 5.2: Cumulative coverage over candidate instances by the most common words to follow candidate instances in the Enhanced Cues Corpus.

5.3 Approach

5.3.1 Preliminaries

The approach discussed in this section will treat sentence-level detection and word-level delineation as separate tasks. Once a sentence is predicted to contain mentioned language, a specific subsequence of words within it can be given a predictive label.

Before these tasks, it was necessary to manually re-label some of the sentences in the Enhanced Cues Corpus. This is because the original labeling procedure had focused upon candidate instances as defined in Section 4.4, but some instances of mentioned language occurred in corpus sentences outside of candidate instances (i.e., outside of stylistic cues or not near any mention words, or both). This meant that the positive and negative labels for candidate instances could not directly apply to their associated sentences, as those labels might underreport mentioned language. Thus, the positive labels for candidate instances were “promoted” to apply to their associated sentences, but each sentence containing a negative candidate instance was re-examined for mentioned language in other places and then labeled accordingly. This led to a refactored corpus of 564 “mention” sentences (containing mentioned language) and 1,350 “non-mention” sentences.

Expectations of the performance of mentioned language detection (and, to a lesser extent, delineation) were tempered by the results of the inter-annotator agreement study discussed in 4.4.5. All four annotators agreed on the presence or non-presence of mentioned language for 76% of the 100 instances, with a Kappa score of 0.74. The Kappa score between the primary annotator and the “majority vote” of the three additional annotators was 0.90. When judging the performance of sentence classifier, an F-measure substantially greater than these Kappa scores would have questionable value. In effect, the classifier would agree with the researcher’s annotations more often than mentioned language is generally agreed upon by well-informed human annotators.

5.3.2 Detection Task

No previous literature was available on using machine learning methods to detect mentioned language in text, and it was necessary to establish baseline levels of performance on the detection task. A matrix of feature sets and classifiers was run on the Enhanced Cues Corpus, using 10-fold cross validation. The following baseline feature sets were constructed, shown below with abbreviations for future brevity:

- stemmed unigrams, i.e., bag of stemmed words (SU)

- unstemmed unigrams (UU)
- stemmed unigrams plus stemmed bigrams (SUSB)
- unstemmed unigrams plus unstemmed bigrams (UUUB)

Classifiers were chosen to reflect a variety of approaches to supervised learning; as implemented in Weka (Hall et al. 2009), these were Naïve Bayes (John and Langley 1995), SMO (Keerthi et al. 2001), J48 (Quinlan 1993), IBk (Aha and Kibler 1991), and Decision Table (Kohavi 1995).

The observations in Section 5.2 on mention words suggested a possible improvement beyond the baseline feature sets. Since a small set of metalinguistically significant words appeared to correlate with mentioned language, a "core mention words" (henceforth abbreviated CMW) approach was taken to feature selection[22]. SU features were ranked by information gain, and all except the top ten features were discarded. The pruning was done using the training set for each of the ten cross-validation folds: that is, the training and testing data for a fold were pruned to the top ten features as ranked from the training instances (with no inspection of the testing instances). The five selected classifiers were then applied to the data.

5.3.3 Delineation Task

The delineation task focused on identifying the sequence of words in a sentence most likely to be mentioned language. To simplify the problem, an instance of mentioned language was assumed to be a sequence of *consecutive* words, with forgiveness during evaluation for the inclusion of certain frequent non-mentioned words. These were generally words like *and* in the sentence

[22] Some attempts were made to identify words *inversely* correlated with mentioned language, but as intuition might predict, those attempts were unsuccessful.

(1) Snow and hail refer to frozen precipitation.

or *the* in cases when it was debatable whether the determiner ought to be included. Metalanguage played a key role in this task as well, as metalinguistic terms anchored the procedures used to delimit mentioned language. During the Combined Cues Study and the Enhanced Cues study, two very frequent relationships were observed between mention words and mentioned language. The first relationship was *noun apposition*, in constructions such those in bold in (2) and (3) below.

(2) The **term cracker** was used in Elizabethan times to describe braggarts.

(3) Confire comes from the Latin **word conficere**.

The second relationship was *mentioned language in a semantic role for a mention verb*. Examples of this are italicized in (4) and (5) below.

(4) Hence, this is sometimes **called the alpha profile**.

(5) **Speil** sometimes **refers** to an informal curling game.

Notably these patterns do not guarantee mentioned language, as illustrated in (6) and (7) below, but they provide a starting point for delineation.

(6) Hence, I called the inspector general.

(7) Jason sometimes refers to the manual for instructions.

For the delineation task, case studies were performed on three of the most frequently occurring *mention patterns* observed in the corpora. The first two were noun apposition with *term* and *word*, as illustrated in (2) and (3), and the third was the appearance of a specific semantic role attached to the verb *call*, as illustrated in (4). *Term*, *word*, and *call* were the most common words to occur in proximity to candidate instances, and all of them had precisions of at least 80% for retrieving mentioned language. These case studies used mention patterns manually identified from the entire corpus and then tested on all relevant sentences (i.e., those containing the relevant mention words); as such, this was not a true experiment.

A common procedure was followed for *term* and *word*. First, sentences containing the mention word were extracted from the Enhanced Cues Corpus; there were 91 of these sentences for *term* and 107 for *word*. These sentences were parsed by the Stanford Parser (Klein and Manning 2003), and then processed using TRegex and TSurgeon (Levy and Andrew 2006) to identify instances of noun phrases in apposition with the mention word. TRegex is a tool for regular expression-like searching of parse trees, and TSurgeon makes alterations to parse trees based on TRegex search results. Though an explanation of their syntax is beyond the scope of this chapter, Figure 5.3 below shows the TRegex and TSurgeon search-alteration pair for *term*. The pair for *word* was identical save for the substitution of *word* for *term.*

```
NP < (DT <: /[Tt]he/ $++ (/^N.*/ <: /term[s]?/ $++ /^[JNVAPDR].*/=mw))
relabel mw MW
```

Figure 5.3: TRegex pattern and TSurgeon command for identifying mentioned language in apposition with *term*.

In essence, the search pattern requires *the* to occur at the beginning of a noun phrase, followed by *term*, followed by a constituent to be relabled *MW* by TSurgeon. Zero or more unrelated words could intervene between *the, term*, and the *MW* phrase, as long as all three occurred in order as siblings in the same noun phrase. The *MW* relabeling was not intended to propose mentioned language as a unique part of speech but rather to make the identified phrase apparent in the output.

The verb *call* occurred in 158 sentences in the Enhanced Cues Corpus. The Illinois Semantic Role Labeler (SRL) (Zimak et al. 2011) was used to identify arguments of *call* in and their semantic roles in each of these sentences. Early observations found that SRL generally labeled mentioned language as an attribute of another argument to *call*. This role (“attribute of arg1”, in SRL terms), when attached to *call*, was considered the predictive label for mentioned language. The results of the predicted labels were thus compared to the refactored Enhanced Cues Corpus.

5.4 Results and Discussion

5.4.1 Detection Task

Tables 5.1-5 below show the performances of the selected classifiers on each feature set. Figures shown for precision, recall, and F-score are the arithmetic means of the ten cross-validation runs.

Stemmed Unigrams			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.759	0.630	0.688
SMO	0.739	0.673	0.704
IBk	0.690	0.642	0.664
Decision Table	0.755	0.609	0.673
J48	0.721	0.686	0.702

Table 5.1: Results of classifiers using the SU feature set.

Unstemmed Unigrams			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.753	0.626	0.682
SMO	0.780	0.638	0.701
IBk	0.701	0.598	0.643
Decision Table	0.790	0.575	0.664
J48	0.761	0.639	0.693

Table 5.2: Results of classifiers using the UU feature set.

Stemmed Unigrams Plus Stemmed Bigrams			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.750	0.591	0.659
SMO	0.776	0.688	0.727
IBk	0.683	0.645	0.661
Decision Table	0.752	0.632	0.684
J48	0.735	0.699	0.714

Table 5.3: Results of classifiers using the SUSB feature set.

Unstemmed Unigrams Plus Unstemmed Bigrams			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.760	0.581	0.657
SMO	0.794	0.648	0.712
IBk	0.682	0.575	0.623
Decision Table	0.778	0.575	0.659
J48	0.774	0.650	0.705

Table 5.4: Results of classifiers using the UUUB feature set.

Core Mention Words			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.750	0.602	0.664
SMO	0.754	0.703	0.727
IBk	0.744	0.720	0.731
Decision Table	0.743	0.684	0.711
J48	0.746	0.733	0.739

Table 5.5: Results of classifiers using the CMW feature set.

To determine whether CMW was an improvement over the baseline feature sets, baseline F-scores were compared to CMW F-scores on a by-classifier basis. (For example, SU-Naïve Bayes was compared to CMW-Naïve Bayes, SUSB-SMO was compared to CMW-SMO, etc.) One-tailed T-tests were used to identify statistically significant improvements by CMW over the baselines. The sample sets were the F-scores from cross validation runs, and a 95% confidence level was used for all tests. The cross-validation partitions for SU and CMW were the same, which enabled paired T-tests for those comparisons; for the rest, standard T-tests were used. The results of all the significance tests appear in Table 5.6 below.

As shown, CMW-Naïve Bayes made no significant gains over the baseline feature sets, probably due to violations of the Naïve Bayes assumption. SUSB-SMO, UUUB-Naïve Bayes, and UU-Naïve Bayes all had higher F-scores than their CMW equivalents; however, none of these were statistically significant deficiencies by CMW (again determined by one-tailed standard T-tests). The highest F-score overall was registered by CMW-J48, at 0.739; however, this was a significant improvement over only the unigram feature sets. The second highest performance was registered by

CMW-IBk, close behind at 0.731. This result *was* significantly better than for IBk on all other feature sets, suggesting that either CMW-J48 or CMW-IBk might be the best performer of the 25 feature set-classifier combinations tested. Both of those combinations also have exceptionally well-balanced precision and recall compared to the population at large. Although the differences were sometimes slight, these results seem to suggest that the CMW approach provides an improvement over the baselines. CMW's closest relative among the baselines is SU, and for four of the five classifiers CMW showed a significant advantage over it.

Classifier	**SU**	**UU**	**SUSB**	**UUUB**
Naïve Bayes				
SMO	●			
IBk	●	○	○	○
Decision Table	●	○		○
J48	●	○		

Table 5.6: Matrix of results from statistical significance tests. A dot indicates a statistically significant improvement by CMW over the baseline: filled dots indicate a paired T-test was used, and hollow dots indicate a standard T-test was used.

Examining the chosen CMW feature sets revealed that most of the features were "mention words": nearly all of them were metalinguistically significant and appeared in Figure 5.1 or Figure 5.2. Specifically, the following nine words appeared in all the feature sets for the ten folds of CMW: *name, word, call, term, mean, refer, use, derive,* and *Latin*. The two last words are perhaps artifacts of the encyclopedic nature of the source text, but the rest appear to generalize more easily. Given the juxtaposition of these findings and the composition of the Combined Cues Corpus (as discussed in Section 4.4.4), the researcher would hypothesize that the mere presence of a core mention word in a sentence is often sufficient to make a positive prediction, although certain combinations of mention words are likely to be better predictors than solitary occurrences. Future research using additional text sources will be necessary to fully establish whether the CMW approach (as well as the specific metalinguistic terms listed above) are widely applicable. However, it also appears that between 20% and 30% of instances of mentioned language remain stubbornly difficult to identify using unigram and bigram-based features alone. More sophisticated features that exploit sentence semantics might be necessary to improve upon the performances shown in this chapter.

Finally, the best CMW performances approach the level of the four-annotator Kappa score from Section 4.4.5. Although this is an indication of some success, the researcher believes that the higher two-party Kappa score of 0.90 remains a meaningful goal for future research efforts.

5.4.2 Delineation Task

In light of the results of the detection task, all sentences that contained *term*, *word*, and *call* were processed by the respective delineation procedures for each of those words. This was done to see if the procedures could provide more fine-grained discrimination between mention and non-mention sentences, effectively taking on the detection task as well. Table 5.7 below shows the results of the delineation procedures in two forms:

- Pattern Application: Here, precision and recall have nearly the same meaning as for the detection task. If a mention pattern was applicable to an instance of the appropriate term in a sentence, it was considered a positive prediction.
- Label Scope: These numbers were collected for true positive predictions. It was determined whether the labeled sequence of words was exactly correct (save for the "forgiveness" described in 5.3.3), too broad (covering extra words), or not broad enough.

Word	**Pattern Application**			**Label Scope**		
	Precision	**Recall**	**F1**	**Overlabeled**	**Underlabeled**	**Exact**
term (n)	1.0	0.89	0.90	0	2	57
word (n)	1.0	0.94	0.97	3	4	57
call (v)	0.87	0.76	0.81	16	1	68

Table 5.7: Performance measures of the mention patterns for the three selected metalinguistic terms.

For the two nouns, *all* applications of their mention patterns were correct, resulting in perfect precision. Just a few instances of *term* and *word* were falsely rejected by their patterns, leading to less than perfect (albeit still high) recall scores. Precision and recall for *call* were moderately high but less exemplary than for the nouns; this seemed due in part to the difficulties of accurate semantic role labeling (Màrquez et al. 2008), as the output of SRL often was flawed.

Overlabeling and underlabeling were only slight problems for *term* and *word*, but overlabeling occurred more frequently for c*all*. Often, the label would "spill" far past the actual end of mentioned language, due to the boundaries of the semantic role in the SRL output. For example, the entire phrase in bold in (8) below was erroneously predicted to be mentioned language, instead of simply *snow-eaters*:

(8) Winds of this type are called snow-eaters for their ability to make
snow melt or sublimate rapidly.

It seemed that the syntactic approach to labeling used for *term* and *word* was more reliable than the semantic approach for *call*. Future improvements in semantic role labeling may remedy this gap.

Overall, these results suggest that the procedure used for the delineation of mentioned language can also handle detection of the phenomenon. In general, when one of the selected metalinguistic terms occurred without mentioned language, its mention pattern simply did not apply. Moreover, the procedure for creating more mention patterns is scalable. The appositive approach used for *term* and *word* could be extended to other appropriate metalinguistic nouns (e.g., *name, title*, *phrase*, *symbol*, etc.) with minimal modifications. The semantic role approach is extensible as well, though the relevant roles are likely to differ among metalinguistic verbs. Additionally, other categories of mention patterns are known to exist, such as a category for copular sentences:

(9) *Igloo* is a term for a domed dwelling built out of snow.

An informal perusal of the corpora constructed for this project suggests that the total number of categories is small, and that the greater difficulty lies in correct discrimination between when to apply patterns (thus predicting mentioned language) and when to abstain.

5.5 Conclusions

This chapter described an effort to computationally identify mentioned language in sentences without the use of stylistic cues. A variety of classifier and feature sets combinations were tested against a core mention words approach for detecting mention sentences. This “CMW” approach was shown to perform significantly better than many of the baseline combinations, and some evidence suggests that the approach is superior overall. Delineation of mentioned language was accomplished using cues in syntax and semantic roles. Case studies of delineation were performed on mention words *call*, *term*, and *word*; these showed good performance not only for delineation but for detection as well (albeit given the presence of *call*, *term*, or *word* in each sentence). The detection and delineation methods presented in this chapter are the first steps toward a computationally practical approach to labeling mentioned language in arbitrary text.

Chapter 6: Closing Thoughts

6.1 Discussion

Whereof one cannot speak, one must pass over in silence.

—Ludwig Wittgenstein (1889-1951)

This project has made the first efforts toward computational identification of mentioned language. Like many problems in natural language processing, it is difficult or even impossible to perform perfectly. The variability in how humans interpret mentioned language represents a ceiling to performance, and the elusiveness of the abilities often termed "AI completeness" (Sloman 1993) is perhaps also an obstacle. The accumulated intelligence of a person is a likely requirement to interpret all the signs of the use-mention distinction, in syntax, semantics, pragmatics, context, and world knowledge.

Still, many things now can be said about mentioned language that previously could only be postulated by theory or mere intuition. Recurring, significant patterns have been observed in mentioned language, and laboratory examples of the phenomenon would not have been sufficient to identify those patterns. The hypothesis stated in Section 1.3 has been validated for a subset of varieties of mentioned language, with some promising indications for future efforts. The researcher believes that these are crucial first steps for moving the study of metalanguage and the use-mention distinction from theory to practice.

6.2 Contributions

The following contributions have been made by this dissertation, with the hope that they will be useful for future efforts to study the same phenomenon or to apply this work to related problems:

- The goal of giving conversational adequacy to computers was advanced through enhancements to the ALFRED dialog system. The system is capable of engaging in conversation repair strategies to resolve problems in communication. ALFRED's concept space represents language knowledge in an explicit, rich form that permits uniform reasoning over both language and domains. ALFRED is now capable of storing knowledge obtained from mentioned language, though substantial future work will be necessary to implement transformations from "raw" user utterances into their equivalent concept space representations.
- A framework has been constructed for identifying and analyzing the use-mention distinction. Mentioned language, metalanguage, and quotation have been defined consistently and intuitively, in an effort to surpass limitations and conflicts in the previous work. To operationalize the definition of mentioned language, a rubric was constructed for labeling the phenomenon, and its equivalence to the definition was shown. Many qualities of mentioned language have been gathered into an enumerated list, in order to specify boundaries for this work and for subsequent studies.
- Labeled corpora of mentioned language have been constructed, with the aid of lexical and stylistic cues. Two inter-annotator agreement studies have

quantified the agreement among human readers in labeling the phenomenon. Instances of the phenomenon are delineated in the corpora, enabling analysis of how it relates to sentence structure. The corpora illustrate the varieties of mentioned language, and attempts have been made to categorize the phenomenon, though a truly decisive set of labels seems elusive. The corpora have been preserved in an XML format for future researchers to use.

- Patterns were identified in vocabulary, sentence structure, and semantic roles that enable the detection and delineation of mentioned language. Detection using a “core mention words” approach to feature selection was shown to perform significantly better than baseline approaches for some supervised classifiers. Delineation procedures for some forms of mentioned language were constructed by exploiting relationships between the phenomenon and three common metalinguistic terms. The full breadth of the phenomenon is likely to elude computational tools for the foreseeable future, but some common cues have been shown to perform well.

6.3 Future Work

Following the course set by this project, some future research directions exist across artificial intelligence, computational linguistics, and natural language processing. Several practical applications were described in Section 1.2: lexical semantics tools, conversational agents, the study of trends in language, source attribution, finely tuned language learning materials, and copyediting software. Such applications will be important for merging this project into the larger goals of research on natural language and computing. Additionally, some opportunities exist for further basic research into metalanguage and the use-mention distinction:

- Mentioned language in informal and conversational English will deserve a dedicated study. The researcher expects that many of the lexical cues discussed in this dissertation will remain the same, but cues in utterance structure will be different and perhaps more difficult to detect.
- English is certainly not the only language in which the use-mention distinction occurs, and studying it in other natural languages should be possible using methods similar to those in this dissertation.
- The semantics of the use-mention distinction deserve further examination with regard to how the distinction naturally occurs in language production. Although the semantics of the distinction have received substantial philosophical treatments (discussed in Chapter 3), those studies were influenced only slightly (if at all) by the actual patterns and tendencies that language users exhibit. The findings in this dissertation present an opportunity to remedy that and construct a more “natural” semantics for use and mention.

Appendix A: Annotation Guidelines

Participant Instructions

This task is part of a study of the use-mention distinction in the English language. The purpose of the task is to identify instances of *mentioned language* in sentences from Wikipedia. Your answers will be aggregated and compared with those of other participants to build a corpus of instances that can be examined for recurring patterns in language.

You will be presented with a list of sentences, each containing a phrase between two star symbols (*). You must read each sentence and place it in one of five categories by interpreting the role of the starred phrase. Unclear cases are expected to occur; do your best using the guidelines in these instructions.

Here are the five categories, with positive and negative examples to illustrate them.

Words as Words (W): If, within the context of the sentence, the starred phrase is used to mean the word or phrase itself and not what it usually refers to, then choose this category.

Words as Words	**Not** Words as Words
Cheese is derived from a word in Old English.	*Cheese* is derived from milk.
This kind of drink is called *iced mocha*.	This drink is an *iced mocha*.
The verb *help* has several different senses.	I *help* the committee in several ways.

Names as Names (N): If the sentence directly refers to the starred phrase as a proper name, nickname, or title, then choose this category.

Names or Titles as Themselves	**Not** Names or Titles as Themselves
Spain is the name of a European country.	*Spain* is a European country.
The next book was called *Speaker for the Dead*.	The next book was *Speaker for the Dead*.
The pseudonym *John Doe* has precedent.	He described John Doe as *a bit odd*.

Spelling or Pronunciation (S): If the starred text appears *only* to illustrate spelling, pronunciation, or a character symbol, then choose this category.

Spelling or Pronunciation	**Not** Spelling or Pronunciation
Australians say the city name as *canz*.	Residents of *Cairns* say the city's name oddly.
Garcia spelled his name with *ie* back then.	*Garcia* spelled his name with ie back then.
Plauche is pronounced *PLO-SHAY*.	*Plauche* pronounces her name PLO-SHAY.

Other Mention/Interesting (O): Use this category at your own discretion to highlight sentences that do not fit into the above categories but still seem relevant to the project. Sentences in this category will receive further examination.

Not Mention (X): Use this category if none of the above categories applies.

If a sentence does not fit precisely in one category, or you are uncertain which category to place it in, you may place it in a second category as well. However, you must designate one category as "primary" and the other as "secondary", and you may not place a sentence in more than two categories.

If it helps, you may use this informal test to determine whether a starred phrase is mentioned language: *If I spoke the sentence to someone in person, and I substituted the starred phrase with the word "that" and pointed to the phrase written on a piece of paper, how much would the meaning change?* You can think of this test as separating the phrase from what it might refer to (e.g., the word "car" from an actual car, or the name "Alice" from a person named Alice).

Thank you for your participation.

Appendix B: Pseudocode of the Rubric

Presented here is a pseudocode equivalent of the rubric for mentioned language in Section 2.5.1. As is the case with the rubric, some exceptions to the pseudocode equivalent exist regarding quotation marks and redundant wording; these are examined in detail in Section 2.5. The series of instructions below is intended to be a simple rewording to make the steps of applying the rubric as clear and distinct as possible.

Given S a sentence and X a copy of a linguistic entity in S:

(1) Create X': The phrase "that [item]", where [item] is the appropriate term for linguistic entity X in the context of S.

(2) Create S': Copy S, and replace the occurrence of X with X'.

(3) Create W: the set of truth conditions of S.

(4) Create W': the set of truth conditions of S', given the assumption that X' in S' is understood to refer deictically to X.

(5) Compare W and W'. If they are equal, X is mentioned language in S. Else, X is not mentioned language in S.

To illustrate, here are two examples.

<u>Example 1</u>: Positive Example

S is the sentence *Spain is the name of a European country.*

X is *Spain*.

(1) Create X': *that name*

(2) Create S': *That name is the name of a European country.*

(3) Create W: Stated briefly, *Spain* is the name of a European country.

(4) Create W': Stated briefly, *Spain* is the name of a European country.

(5) Compare W and W': They are equal. *Spain* is mentioned language in S.

Example 2: Negative Example

S is the sentence *Spain is a European country*.

X is *Spain.*

(1) Create X': *that name*

(2) Create S': *That name is a European country.*

(3) Create W: Stated briefly, Spain is a European country.

(4) Create W': Stated briefly, the name *Spain* is a European country.

(5) Compare W and W': They are not equal. *Spain* is not mentioned language in S.

The reader may consult Section 2.5.2 for more examples of the rubric in action.

Appendix C: The Token-Type Distinction and the Definition

The token-type distinction is described in the Stanford Encyclopedia of Philosophy (Wetzel 2011):

> The distinction between a type and its tokens is an ontological one between a general sort of thing and its particular concrete instances…Types are generally said to be abstract and unique; tokens are concrete particulars, composed of ink, pixels of light (or the suitably circumscribed lack thereof) on a computer screen, electronic strings of dots and dashes, smoke signals, hand signals, sound waves, etc.

This distinction provides a mechanism for discussing the abstract properties of specific linguistic entities (i.e., the properties of their types) using concrete instantiations of those entities (the tokens of their types). Such discussion is a primary function of mentioned language: to communicate to the reader or listener generally applicable information about words or phrases, as in (1) below.

(1) "Chair" has five letters.

The information conveyed by (1) will be applicable to any token of *chair*. However, another function of mentioned language is the discussion of tokens, as in (2):

(2) "The" is the first word of this sentence.

The only token of *The* which satisfies the truth conditions of (2) is the token that appears as the first word of (2), and so the type of "The" is not referred to by the sentence. Although it is tempting to describe “The” in (2) as an occurrence of the type *The*, not all occurrences of a type are tokens (Wetzel 2011). Consider that (2) above is a token of a *sentence type*, and that type is composed of a sequence of occurrences of types: *“*, *The, ”*, *is*, *the*, *first*, *word*, etc.

It might be argued that some token-type reference ambiguity is present when mentioning language, as a language user could produce a token either to refer to a property of the token or a property of its type. Sentence (1), for instance, could be further articulated as either (3) or (4) below:

(3) “Chair” the token has five letters.

(4) “Chair” the type has five letters.

Stylistic conventions can be used to distinguish token-reference from type-reference, as done in this appendix except for example sentences (i.e., italics for types, quotation marks for tokens). However, such conventions are often underspecified, since it would be unwise for popular style guides to assume wide familiarity with the token-type distinction. The researcher would hypothesize that language users assume a *maxim* of *quantity of information* (Grice 1975) that biases them toward type-reference while not prohibiting token-reference when the need is apparent.

Bibliography

Aha, David W., and Dennis Kibler. 1991. Instance-based learning algorithms. In *Machine Learning*, 37–66.

American Psychological Association. 2001. *Publication Manual of the American Psychological Association*. 5th ed. Washington, DC: American Psychological Association.

Anderson, Michael L, Yoshi A Okamoto, Darsana Josyula, and Donald Perlis. 2002. “The Use-Mention Distinction and its Importance to HCI.” *In Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialog*: 21--28.

Anderson, Michael L, and Donald Perlis. 2005. “Logic, Self-awareness and Self-improvement: the Metacognitive Loop and the Problem of Brittleness.” *Journal of Logic and Computation* 15 (February): 21–40. doi:10.1093/logcom/exh034.

Anderson, Michael L., Andrew Fister, Bryant Lee, and Danny Wang. 2004. On the frequency and types of meta-language in conversation: a preliminary report. In *14th Annual Conference of the Society for Text and Discourse*. http://www.cs.umd.edu/projects/metalanguage/results/std04_final.pdf.

Anderson, Michael L., Scott Fults, Darsana Josyula, Tim Oates, Donald Perlis, Matthew D. Schmill, Shomir Wilson, and Dean Wright. 2008. “A Self-Help Guide for Autonomous Systems.” *AI Magazine*.

Anderson, Michael L., Matt Schmill, Tim Oates, Donald Perlis, Darsana Josyula, Dean Wright, and Shomir Wilson. 2007. Toward domain-neutral human-level metacognition. In *The 8th International Symposium on Logical Formalizations of Commonsense Reasoning*, 1–6.

Audi, Robert. 1995. *The Cambridge Dictionary of Philosophy*. Cambridge University Press.

Balasuriya, Dominic, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named entity recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources*, 10-18. Suntec, Singapore: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?id=1699767.

Bird, Steven. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, 69–72. COLING-ACL ’06. Stroudsburg, PA, USA: Association for Computational Linguistics.

Blaylock, Nate, William de Beaumont, Lucian Galescu, Hyuckchul Jung, James F. Allen, George Ferguson, and Mary D. Swift. 2010. Learning Collaborative Tasks on Textual User Interfaces. In *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference*.

Bohus, Dan, and Alexander I Rudnicky. 2009. “The RavenClaw dialog management framework: Architecture and systems.” *Computer Speech and Language* 23 (July): 332–361. doi:10.1016/j.csl.2008.10.001.

Cappelen, H, and E Lepore. 1997. “Varieties of quotation.” *Mind* 106 (423) (July 1): 429 -450. doi:10.1093/mind/106.423.429.

Chicago Editorial Staff. 2010. *The Chicago Manual of Style*. 16th ed. University of Chicago Press.

Christensen, Niels Egmont. 1967. “The Alleged Distinction between Use and Mention.” *The Philosophical Review* 76 (3): pp. 358-367.

Clark, Herbert H., and Edward F. Schaefer. 1989. “Contributing to Discourse.” *Cognitive Science* 13 (2): 259-294.

Cohen, Jacob. 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement* 20 (1): 37.

Cram, David F. 1978. “The syntax of direct quotation.” *Cahiers de Lexicologie* 33 (2): 41-52.

Davidson, Donald. 1968. “On Saying That.” *Synthese* 19 (1-2): 130–46.

———. 1979. “Quotation.” *Theory and Decision* 11 (1) (March): 27-40. doi:10.1007/BF00126690.

Dinesh, Nikhil, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, 29–36. Ann Arbor, Michigan: Association for Computational Linguistics, June.

Eskenazi, Maxine, Alan W. Black, Antoine Raux, and Brian Langner. 2008. Let’s Go Lab: a platform for evaluation of spoken dialog systems with real world users. In *Interspeech*. Brisbane, Australia. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.2108.

Fellbaum, Christiane. 1998. *WordNet: an electronic lexical database*. Cambridge: MIT Press.

García-Carpintero, Manuel. 2004. “The deferred ostension theory of quotation.” *Noûs* 38 (4): 674–692.

Geach, P. T. 1950. “On Names of Expressions.” *Mind* 59 (235): 388.

Goldstein, Laurence. 1984. “Quotation of types and types of quotation.” *Analysis* 44: 1-6.

Grice, Paul. 1975. Logic and conversation. In *Syntax and semantics*. Vol. 3. New York: Academic Press.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. “The WEKA data mining software: an update.” *ACM SIGKDD Explorations Newsletter* 11 (November): 10–18. doi:10.1145/1656274.1656278.

Havasi, Catherine, Robert Speer, and Jason Alonso. 2007. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Recent Advances in Natural Language Processing*. Borovets, Bulgaria, September.

Honnibal, Matthew, Joel Nothman, and James R. Curran. 2009. Evaluating a statistical CCG parser on Wikipedia. In *Proceedings of the 2009 Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources*, 38-41. Suntec, Singapore: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?id=1699765.1699771.

Hu, Guangwei. 2010. “A place for metalanguage in the L2 classroom.” *ELT Journal*. doi:10.1093/elt/ccq037. http://eltj.oxfordjournals.org/content/early/2010/06/21/elt.ccq037.abstract.

John, George, and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 338–345. Morgan Kaufmann.

Jorgensen, Julia, George A. Miller, and Dan Sperber. 1984. "Test of the mention theory of irony." *Journal of Experimental Psychology: General* 113 (1) (March): 112-120.

Josyula, Darsana. 2005. A unified theory of acting and agency for a universal interfacing agent. College Park, MD, USA: University of Maryland at College Park.

Josyula, Darsana, Michael L Anderson, and Donald Perlis. 2003. Towards domain-independent, task-oriented, conversational adequacy. In *Proceedings of the 18th International Joint Conference on Artificial intelligence*, 1637–1638. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://portal.acm.org/citation.cfm?id=1630659.1630950.

Josyula, Darsana, Scott Fults, Michael L. Anderson, Shomir Wilson, and Donald Perlis. 2007. Application of MCL in a dialog agent. In *Papers from the Third Language and Technology Conference*. Poznań, Poland.

Keerthi, S. S, S. K Shevade, C. Bhattacharyya, and K. R. K Murthy. 2001. "Improvements to Platt's SMO Algorithm for SVM Classifier Design." *Neural Comput.* 13 (March): 637–649. doi:10.1162/089976601300014493.

Kirkham, Richard. 1995. *Theories of Truth : A Critical Introduction*. 1st ed. Cambridge, MA: MIT Press.

Kiss, Tibor, and Jan Strunk. 2011. "Unsupervised Multilingual Sentence Boundary Detection." *Computational Linguistics* 32 (4) (January 31): 485-525. doi:10.1162/coli.2006.32.4.485.

Klein, Dan, and Christopher D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems*. Vol. 15. The Stanford Natural Language Processing Group. http://www-nlp.stanford.edu/ manning/papers/lex-parser.pdf.

Kohavi, Ron. 1995. The Power of Decision Tables. In *Proceedings of the European Conference on Machine Learning*, 174–189. Springer Verlag.

Kumar, Rohit, and Carolyn P Rosé. 2009. Building conversational agents with Basilica. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session*, 5–8. NAACL-Demonstrations '09. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lester, James, Karl Branting, and Bradford Mott. 2004. Conversational Agents. In *The Practical Handbook of Internet Computing*, ed. Munidar Singh. Chapman & Hall/CRC Press.

Levy, Roger, and Galen Andrew. 2006. TRegex and TSurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation*.

Lin, Ziheng, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1:343–

351. EMNLP ’09. Stroudsburg, PA, USA: Association for Computational Linguistics.

Litman, Diane J, and Shimei Pan. 2002. “Designing and Evaluating an Adaptive Spoken Dialogue System.” *User Modeling and User-Adapted Interaction* 12 (March): 111–137. doi:10.1023/A:1015036910358.

Liu, H., and P. Singh. 2004. “ConceptNet — A Practical Commonsense Reasoning Tool-Kit.” *BT Technology Journal* 22 (October): 211–226. doi:10.1023/B:BTTJ.0000047600.45421.6d.

Lynch, Michael. 2001. *The Nature of Truth : Classic and Contemporary Perspectives*. Cambridge, MA: MIT Press.

Maier, Emar. 2007. Mixed quotation: between use and mention. In *Logic and Engineering of Natural Language Semantics Workshop*. Miyazaki. http://ncs.ruhosting.nl/emar/em_lenls_quot.pdf.

Marcus, Mitchell P, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. “Building a large annotated corpus of English: the penn treebank.” *Comput. Linguist.* 19 (June): 313–330.

Màrquez, Lluís, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. 2008. “Semantic role labeling: an introduction to the special issue.” *Computational Linguistics* 34 (June): 145–159. doi:http://dx.doi.org/10.1162/coli.2008.34.2.145.

Miller, Michael J. 1993. A view of one’s past and other aspects of reasoned change in belief. PhD thesis, University of Maryland at College Park. http://portal.acm.org/citation.cfm?id=164828.

Nadeau, David, and Satoshi Sekine. 2007. “A survey of named entity recognition and classification.” *Lingvisticae Investigationes* 30 (1) (January): 3-26.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. “The Proposition Bank: An Annotated Corpus of Semantic Roles.” *Computational Linguistics* 31 (1) (March): 71-106. doi:10.1162/0891201053630264.

Partee, Barbara. 1973. The syntax and semantics of quotation. In *Festschrift for Morris Halle*, ed. Stephen Anderson and Paul Kiparsky. New York: Holt, Rinehart, Winston.

Perlis, Donald, Khemdut Purang, and Carl Andersen. 1998. “Conversational adequacy: mistakes are the essence.” *International Journal of Human-Computer Studies* 48 (5): 553-575.

Quine, W. V. O. 1940. *Mathematical logic*. Cambridge, MA: Harvard University Press.

Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. San Mateo Calif.: Morgan Kaufmann Publishers.

Raux, Antoine, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let’s Go public! Taking a spoken dialog system to the real world. In *Proceedings. of Interspeech 2005*.

Reimer, Marga. 1996. “Quotation Marks: Demonstratives or Demonstrations?” *Analysis* 56 (3): 131–141.

Richard, Mark. 1986. “Quotation, Grammar, and Opacity.” *Linguistics and Philosophy* 9 (3): 383-403.

Riloff, Ellen, Janyce Wiebe, and William Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence*, 3:1106–1111. AAAI Press. http://portal.acm.org/citation.cfm?id=1619499.1619511.

Russell, Stuart, and Peter Norwig. 1995. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, N.J.: Prentice Hall.

Saka, Paul. 1998. "Quotation and the use-mention distinction." *Mind* 107 (425) (January 1): 113 -135. doi:10.1093/mind/107.425.113.

———. 2003. "Quotational Constructions." *Belgian Journal of Linguistics* 17 (1) (January). doi:10.1075/bjl.17.11sak.

Schmill, Matthew D., Darsana Josyula, Michael L. Anderson, Shomir Wilson, Tim Oates, Donald Perlis, Dean Wright, and Scott Fults. 2007. Ontologies for reasoning about failures in AI systems. In *In Proceedings from the Workshop on Metareasoning in Agent Based Systems at the Sixth International Joint Conference on Autonomous Agents and Multiagent Sytems*.

Shapiro, Stuart C., and Michael Kandefer. 2005. A SNePS Approach to The Wumpus World Agent or Cassie Meets the Wumpus. In *IJCAI-05 Workshop on Nonmonotonic Reasoning, Action, and Change: Working Notes*, ed. Leora Morgenstern and Maurice Pagnucco, 96–103. Edinburgh, Scotland: IJCAII.

Shapiro, Stuart, William Rapaport, Michael Kandefer, Frances Johnson, and Albert Goldfain. 2007. "Metacognition in SNePS." *AI Magazine* 28 (1): 17-31.

Sloman, Aaron. 1993. Prospects for AI as the general science of intelligence. In *Prospects for Artificial Intelligence*, 1-10. IOS Press.

Sperber, Dan, and Deirdre Wilson. 1981. Irony and the Use-Mention Distinction. In *Radical Pragmatics*, 295-318. New York.

Strunk, Jr., and E. B. White. 1979. *The Elements of Style*. Third. Macmillan.

Tarski, Alfred. 1933. The concept of truth in formalized languages. In *Logic, Semantics, Mathematics*, ed. J. H. Woodger. Oxford: Oxford University Press.

Traum, David R, Lenhart K Schubert, Massimo Poesio, Nathaniel G Martin, Marc N Light, Chung H Hwang, Peter A Heeman, George M Ferguson, and James F Allen. 1996. *Knowledge Representation in the TRAINS-93 Conversation System*. Rochester, NY, USA: University of Rochester.

Washington, Corey. 1992. "The identity theory of quotation." *Journal of Philosophy* 89 (11): 582-605.

Wetzel, Linda. 2011. Types and Tokens. Ed. Edward N. Zalta. *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/spr2011/entries/types-tokens/.

Ytrestøl, Gisle, Dan Flickinger, and Stephan Oepen. 2009. Extracting and annotating Wikipedia sub-domains. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories*.

Zesch, Torsten, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of LREC '08*. Marrakech, Morocco: European Language Resources Association.

Zimak, Dav, Ming-Wei Chang, Vasin Punyakanok, and Wen-tau Yih. 2011. *Illinois Semantic Role Labeler*. University of Illinois at Urbana-Champaign: Cognitive Computation Group.