Spring 2015

# Creating, testing and implementing a method for retrieving conversational inference with ontological semantics and defaults

Tatiana R. Ringenberg
*Purdue University*

*Graduate School Form 30*
*Updated 1/15/2015*

**PURDUE UNIVERSITY**
**GRADUATE SCHOOL**
**Thesis/Dissertation Acceptance**

This is to certify that the thesis/dissertation prepared

By Tatiana R Ringenberg

Entitled

CREATING, TESTING AND IMPLEMENTING A METHOD FOR RETRIEVING CONVERSATIONAL INFERENCE WITH ONTOLOGICAL SEMANTICS AND DEFAULTS

For the degree of Master of Science

Is approved by the final examining committee:

Julia Taylor
Chair

Victor Raskin

John Springer

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Julia Taylor

Approved by: Jeffrey Whitten 4/28/2015
Head of the Departmental Graduate Program Date

# CREATING, TESTING AND IMPLEMENTING A METHOD FOR RETRIEVING CONVERSATIONAL INFERENCE WITH ONTOLOGICAL SEMANTICS AND DEFAULTS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Tatiana R Ringenberg

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2015

Purdue University

West Lafayette, Indiana

For my parents Paul and Carolea Ringenberg.

# ACKNOWLEDGEMENTS

I am very thankful to my major Professor, Dr. Julia Taylor, for her assistance, support and patience. I am also thankful to Dr. Victor Raskin and Dr. John Springer for their guidance and suggestions.

I am thankful to Trenton Epperson for helping me troubleshoot issues with many of the tools discussed in this thesis. I am also thankful to Lauren Stuart for helping me understand dependencies and dependency parsers.  I am also thankful to Nicholas Valentin for assisting me with identifying formatting mistakes.

# TABLE OF CONTENTS

Page

Page

APPENDICES

# GLOSSARY

Acquisition – In this research, acquisition refers to the process of gathering and inserting information into an Ontology, Lexicon or other repository within Ontological Semantics Technology.

Concept – A concept is a representation of an entity or idea.

Corpus – A corpus is a collection of texts generally referred to in language or literary studies (Kilgarriff & Grefenstette, 2003).

Default – A default is any information, associated with an event, which a communicator finds to be too trivial to state.

Maxim of Quantity – A speaker provides as much information as is required. A listener expects unambiguous and concise information (Atifi, Mandelcwajg & Marcoccia 2011).

Mutual Knowledge – "type of knowledge which two (or more) persons hold to be common with 100% certainty" (Lee 2011).

Ontology - constructed world model based on human perception (Nirenburg & Raskin, 2001).

Ontological Semantics - Ontological semantics is a theory of meaning in natural language and an approach to NLP which uses a constructed world model, or ontology, as the central resource for extracting and representing meaning of natural language texts, reasoning about knowledge derived from texts as well as generating natural language texts based on representations of their meaning (Nirenburg & Raskin, 2001).

Ontological Semantic Technology – Ontological Semantics Technology (OST) is an implementation of Ontological Semantics used to detect meaning in text. OST uses several tools including but not limited to a language-independent lexicon and ontology, a Text Meaning Representation generator and an InfoBase (Taylor, Hempelmann & Raskin, 2010).

Property – A meaningful representation of the relation between concepts.

Referring Expression – A referring expression is a noun phrase that is used to identify a unique object (Van der Sluis, Luz, Breitfuß, Ishizuka, Prendinger 2012).

Typed Dependency – Typed dependency is an easy and straightforward way of representing the grammatical relations between words. A typed dependency is an attempt at providing semantically contentful information for text (De Marneffe & Manning, 2008).

# ABSTRACT

Ringenberg, Tatiana. M.S., Purdue University, May 2015. Creating, Testing and Implementing a Method for Retrieving Conversational Inference with Ontological Semantics and Defaults. Major Professor: Julia Taylor.

Conversational inference refers to that information which is assumed to be understood by both speaker and listener in conversation. With conversational inference, a speaker makes the assumption that what is being omitted from the conversation is already known by the listener. In return, a listener assumes that the information that the listener perceives to be omitted is the same as what the speaker believes to be omitted.

Ontological Semantic defaults represent the information which is implied in a single event. Defaults are typically excluded from conversation unless new information is being presented or the speaker is purposefully emphasizing the default for some reason.

Little research has been done in the area of defaults. This thesis expands the research on defaults through the implementation and adjustment of an algorithm for default detection.

The investigation into default detection is broken into two phases. In the first phase, the original algorithm for default detection is implemented. This algorithm involves pulling defaults based on adjectival modifiers to an object associated with an event. Phase 2 expands the algorithm from Phase 1 to include several additional modifiers. The algorithm from Phase 2 is found to be more effective than that in Phase 1.

# CHAPTER 1. INTRODUCTION

When two people hold a conversation the meaning of each individual word is rarely taken as literal. Much of understanding in conversation is implied and gathered over time between the participants or generally understood by the population. For instance, when people are having a conversation about driving they very rarely mention that what they are driving is a car. The car is not stated but implied within the event. Within Ontological Semantics, the information that is not considered, by the speaker, to be important enough to mention in a conversation is referred to as a default. It is information that is not salient in the mind of the speaker and thus is not mentioned directly but is assumed to be understood.

Though significant work has been done on automatic acquisition for ontologies and also on automatically and manually building knowledge bases, very little work has been done on using semantics to pull implied, unstated information based on textual inference. Most work that exists is in the area of machine learning. Though effective, this approach does not take into account semantics. Some approaches do use semantics, however they are largely based on statistical methods and word-relatedness metrics.

Ontological Semantics is a framework for representing meaning, to a computer, the way a human being might. In order to represent the world, concepts which represent a single idea or entity are used. A large set of domain-restricted properties is used to both define and connect concepts together in a way that loosely resembles human logic. These concepts and properties are used to represent meaning of not just individual words but sentences, phrases and larger texts. The rich set of properties associated with Ontological Semantics makes it an ideal choice for the collection of defaults for both large populations and individuals.

The question the researcher sought to answer in this research was: Is the proposed algorithm capable of pulling candidate defaults to increase a computer's understanding of unspoken meaning in text?

The goal of the researcher was to create an algorithm that narrows the range of potential values of a default for a population.

## 1.1 Scope

The topic of defaults is extremely broad. As such, this research focused on identifying candidates for defaults within verb phrases only. Specifically, this research focused on identifying nouns that are potential defaults for a particular verb event. Candidacy for a default was determined by modification of the noun associated with a given verb event.

The researcher only pulled those nouns identified with modifiers in Stanford Collapsed Dependencies. A list of the specific modifiers used shall be presented in Chapter 3.

The method of pulling candidate defaults from verb phrases was loosely based upon the WD-Inference algorithm described in 2010 by Taylor, Raskin, Hempelmann and Attardo.

The ultimate goal of this research was to create and improve algorithms which would assist Ontological Semantic Technology in understanding unspoken meaning.

Verb phrases were chosen because verbs are generally linked to events. However, it is important to note that nouns can also be events. Noun events were outside of the scope of this research. The resources that were available to the researcher were better geared towards verb events than noun events.

Defaults outside of direct and indirect objects were also outside of the scope of this research. As the topic of defaults was so large and the available literature was so small, it was important to keep the investigation restricted to just direct and indirect object defaults.

## 1.2 Significance

A large portion of meaning and understanding comes from unstated mutual knowledge between a speaker and a listener. When a speaker communicates to a listener much of that communication is assumed by the speaker to be already understood by the listener. In exchange, the listener assumes that the speaker understands the information which was not stated in the same way that they do. The information that is not stated is frequently referred to as mutual background knowledge as it is typically assumed to be common to both parties.

However, mutual knowledge is not always common to two parties. Background knowledge can vary from person to person due to differences in language, life experience and culture. As speakers, we are often able to resolve these differences in knowledge by adjusting the level of information we choose to share. In the case of text it is often more difficult to determine that which is implied and to adjust accordingly.

Defaults, within Ontological Semantics, refer to that information which is not stated by the speaker but is assumed to be understood and trivial. Several types of defaults have been identified in the past: basic default fillers, direct object defaults, script defaults and semantic ellipsis. This research specifically focused on direct object defaults.

Direct object defaults were examined as a first step towards automatic acquisition of unspoken information. Availability of unspoken information has significant implications in many fields. Specifically, the researcher saw the value of defaults in health care, Knowledge Representation and Information Security.

Doctor-patient communication is a significant problem in health care. Misunderstandings between healthcare providers and patients potentially lead to misdiagnosis and prolonged illness. As of late, more attention has been placed on the cultural competence of doctors (Paternotte 2015; Teal & Street 2009). Researchers have found that many factors affect the ability to communicate with a physician including cultural background and even gender (Bradley, Sparks & Nesdale 2001; Kule 2012). The detection of defaults has the potential to help understand that which is assumed by both patients and doctors alike. The researcher believed that defaults could be used to find disparities between individuals and potentially populations; defaults could help to bridge the knowledge gap from both sides.

Representing knowledge to a computer is a significant challenge. Currently, most research is focused on statistically examining the text that is stated. Again, this tends to be done with a focus primarily on statistics as opposed to semantics. Default detection added to this space by looking between the lines of a given text and doing it semantically. As neither of these are predominant in computational research, this research on default detection helped to broaden and enrich the space that is still predominantly non-semantic in nature.

Raskin, Taylor and Hempelmann discussed the potential for defaults within Information Security and specifically for insider threat detection (2010). Insider threat has many definitions but essentially refers to "a breach of trust by people within an organization or system" (Bishop, Engle, Peisert, Whalen, Gates, Probst & Somayaji, 2008). Defaults appeared to the researcher to be particularly useful for the detection of lies and unintentional inference which pair well with insider threat detection. As lying is not a form of bonafide communication, a person will generally violate defaults in some way when they lie. Whether intentional or unintentional, default violations are able to be identified if the defaults of the individual are known.

## 1.3 Assumptions

The assumptions for this research included:

- All verbs were representative of events.
- Stanford Parser produced accurate dependencies for sentences.
- Stanford Parser could identify and tokenize sentences correctly.
- The events used in this research were independent.
- Some sentences didn't contain relevant verb-noun and verb-adjective-noun combinations.
- Humans did not communicate everything they meant in conversation explicitly.

- Brown Corpus was accurately tagged.

## 1.4 Limitations

The following were limitations of the research:

- The researcher served as a decision maker in the process.
- The researcher determined the effectiveness of the algorithm.
- Verb events were included in this work.
- The researcher looked into the top 200 verbs from the Brown Corpus.
- The researcher used a selection of documents from Wikipedia to confirm the effectiveness of the algorithm.
- The researcher provided evidence of default inference.
- The researcher examined noun arguments of verb events.
- The researcher analyzed all modifications to nouns associated with verb events except for determiners.
- Duplicate documents were not removed from either dataset.
- Duplicate sentences were not removed from either dataset.

## 1.5 Delimitations

The following were delimitations of the research:

- Noun events were not included in this thesis.
- Intent was not analyzed in this research. Thus intentional and unintentional inferences were both examined and no separated.
- Verbs outside of the 200 most frequent verbs in Brown Corpus were not analyzed.

- Determiners associated with nouns were not examined in this research. The way determiners affect default-ness is still unknown at this time and requires further examination.

## 1.6 Chapter Summary

In Chapter 1, the researcher outlines the necessity for the detection of mutual knowledge and defaults in text.  The scope, significance, assumptions, limitations and delimitations of this research are then given.  In Chapter 2, the researcher will discuss the relevant literature associated with mutual knowledge and defaults.

# CHAPTER 2. LITERATURE REVIEW

In Chapter 2, previous literature available on the concepts of conversational inference and implied meaning is presented. The researcher begins with a discussion of the history of the field. The need for mutual knowledge implementations, current computational solutions touching on mutual and background knowledge and a summary of Ontological Semantics and Ontological Semantics Technology (OST) are also discussed.

## 2.1 Literature Review Background

Mutual knowledge is a topic most frequently covered in the areas of psychology and linguistics. From what the researcher has found, the problem of mutual knowledge is primarily identified and formalized in the 1970s and 1980s. Research on mutual knowledge has continued well into the 2010s but appears to branch from the ideas originally established in the 1970s and 1980s.

To the knowledge of the researcher, little work had been done on applying the principles of mutual knowledge to computation. This review is meant to provide a big picture of the mutual knowledge and implied meaning domain to show where the researcher's study fit into the research community.

## 2.2 Mutual Knowledge and Conversational Inference

Mutual knowledge is a topic most often discussed in pragmatics. Mutual knowledge “assumes that listeners use the knowledge and beliefs they share with speakers in the process of interpreting utterances” (Gibbs, 1987, p. 562). The following sentence demonstrates a situation in which mutual knowledge is required in order to create understanding between a speaker and a listener:

Speaker: Would you like to see a movie tonight?

Listener: I have class.

According to Grice, when humans speak to one another it does not “consist of a succession of disconnected remarks, and would not be rational if they did” (Grice 2013, p. 49). If one takes the literal meaning of the above conversation, the listener’s reaction does not make sense with the conversation. The listener’s response is a non sequitur that does not follow the flow from the speaker. Thus, if a human took this exchange literally, Grice would be correct and it would be seen as an irrational exchange. However humans can understand that the listener’s comment implies that the listener will not be able to go to a movie because they will be in a class tonight.

Most research on the topic of mutual knowledge and conversational inference agrees that a sharing of backgrounds and assumptions takes place in an example such as that above. However, there have been several different views on the usefulness of mutual knowledge in conversation. There have also been several different interpretations of how mutual knowledge is conveyed and exchanged.

The work of H.P. Grice laid the foundation for the modern research that is done on mutual knowledge today. To explain the unspoken connections that people draw during conversation, Grice created the Cooperative Principle. The Cooperative Principle is a series of maxims that outline the rules by which humans seek to communicate information. In the Cooperative Principle, both speakers and listeners are said to make as much of a contribution to the conversation as is required. Both parties shall also make the contribution of information when it is necessary within the conversation. Lastly, both parties shall make contributions that add information to the current topic at hand (Sperber & Wilson, 2006). Grice believed that these maxims were guidelines that people strive towards.

Though Grice did lay the foundation for research in this area, there is much debate over the usefulness of some of his maxims. Researchers have pointed out that Grice’s maxims are extremely vague (Sperber & Wilson, 2006). The terms “relevance” and “clarity” were used to describe the guidelines humans use to communicate information. However, relevance and clarity were not defined. Grice himself acknowledged the difficulties in defining relevance in a true conversation where topics are frequently changing. Instead of addressing the topic he left it to future research (Grice, 2013). For the purposes of computational research, the definition of relevance is extremely important.

Wilson and Sperber (1994) addressed the vagueness of the wording of Grice’s maxims and Cooperative Principle using Relevance Theory. Similar to Grice, Wilson and Sperber (2002) believed that there was not always a comprehendible, literal meaning to a sentence and even when a literal meaning existed it was often not the meaning that was intended by the speaker. However Wilson and Sperber also believed that Grice’s maxims were not tied closely enough with human cognition. Wilson and Sperber claimed that humans focused most on what was relevant in a conversation. As such, humans tend to both state and interpret things in a way that maximizes relevance. This was the basis for the Relevance Theory (Sperber & Wilson, 1994). Sperber and Wilson (p. 44) went on to say that “the greater the contextual effects, the greater the relevance”. The authors also generally noted that interpretations that are easier to make are also higher in relevance.

When one looks closely at Relevance Theory it is clear that it did not negate Grice’s Cooperative Principle but added to it. Consider the statement that greater contextual evidence yields greater relevance. Essentially this statement means that the closer a speaker is to speaking about the topic at hand, the more salient it will be to both the speaker and the listener. This is consistent with Grice’s view that contributions to the conversation should be made on the topic at hand. The statement that easier to understand statements are more relevant is also consistent with Grice. Consider the example of the speaker and listener discussing the movie. The literal explanation of the sentence requires significantly more thought than the implied meaning. The difference between the Grice principle and the Sperber and Wilson theory was the emphasis on relevance as a cognitive process.

Other work in pragmatics that resulted from the work of Grice included referring expressions. Referring expressions are noun phrases that are used to uniquely identify an object. Referring expressions are not just limited to words but often include pointing as well (Beun & Cremers, 1998). The concept of referring expressions was based loosely on Grice’s maxims but went further to define utterances as having goals. It is believed by researchers of referring expressions that we include information about objects as well as specific descriptors for that object because we want to provide additional detail towards some larger goal (Appelt, 1985). Researchers who have worked with referring expressions acknowledge that, often, the information human beings include in their exchanges is redundant and seemingly unnecessary because of these goals (Dale & Reiter, 1995). This redundancy is a violation of Grice's Maxims but yet still represents human communication.

Although not all pragmatists have agreed on how mutual knowledge is exchanged, it is clear that there has been an acknowledgement of common knowledge exchange between speakers and listeners. Some have referred to mutual knowledge exchange as a formal, infinite process (Gibbs, 1987). Others have believed that assumptions made between speaker and listener allow humans to skip the infinite exchange in order to draw conclusions more easily (Clark & Marshall, 1981).

Likely the closest research to the topic of this thesis, Ontological Semantic defaults, has been that of referring expressions. Research in referring expressions has focused on identifying how and when descriptive information is added to an object. Prince (1981) has even acknowledged that some objects are situationally evoked and are therefore a part of a mutually shared context. Goodman (1986) similarly stated that referring expressions should be known to both the speaker and hearer. This was the most default-like observation the researcher had found.

## 2.3 The Need for Mutual Knowledge Implementation

The general acknowledgement of the existence of mutual knowledge exchange in combination with varying views of how this information is obtained demonstrated a need, to the researcher, for research on computational implementation. In order to truly understand human conversation, the researcher believed that a computer would need to obtain background information in some way.

The need for this type of technology has been cited within research on digital collaboration. According to Krauss and Fussell (1990), people have been asking how to address common background knowledge in a global workforce for years. As technology has continued to improve over the years, the number of virtual teams within a company has also increased. Global teams with members all over the have become more common in industry. As such, it has becoming more important to recognize and address cultural differences that can have an effect on the workforce. It was the researcher's belief that mutual knowledge plays a large role in this.

Mutual knowledge refers to shared background and assumptions between a speaker and a listener. If a speaker and a listener do not have this shared background, there is room for error. The consequences of such errors were discussed in detail by Cramton (2001). Cramton noted that some of the consequences that result from inconsistencies in background knowledge include the hesitation of individuals on a team to mention relevant and unique information and the rapid deterioration of working relationships. Both of these observations could have a negative impact on the quality of company projects and on company culture in general.

The researcher believed that this situation could come into play particularly in communication media with unstructured text such as in email, chat or texts. This is due to the lack of visual and vocal cues that would be present in face-to-face contact or even in teleconferencing. Because these barriers exist, it would be extremely useful to develop a tool that could identify this background knowledge for each individual in a conversation.

Another area in which mutual knowledge transfer was cited as important was information security. Specifically within information security, mutual knowledge had been discussed for the purposes of identifying “insider threat”. Insider threat has many definitions but according to Magklaras, Furnell and Brooke (2006) insiders are defined as those who have legitimate access to a company’s IT infrastructure. Insider threat refers to those who misuse that access whether intentionally or unintentionally.

It has been the belief of some that background knowledge, explicit references to background knowledge, and the inclusion of novel information in reference to background knowledge could be used to infer information about an insider who is giving away some piece of information (Raskin, Taylor & Hempelmann, 2010). The example given by Raskin et al. was that of a person saying that the person’s boss asked if they would be willing to fly coach to Germany. It was implied in the Germany sentence that the speaker didn’t frequently fly coach. As such, conclusions could have been drawn about the state of the company. However, in order to pull that knowledge from the text, one would have to notice that the speaker had explicitly stated that they were flying coach. As humans tend to not mention things that are not relevant to the topic, this means that the speaker flying coach was useful information. The authors (Raskin, et al., 2010) argued that this type of information could be pulled from text if the meaning of the sentences was mapped. The paper went on to consider an algorithm for doing so.

## 2.4 Computational Solutions

It is important to make the distinction between pragmatic and semantic background knowledge and background knowledge as it is currently defined and used today. Background and mutual knowledge, for the purposes of this research, referred to common, relevant knowledge as it was outlined by Grice and other pragmatists. The researcher noticed that, within computer science, background knowledge is a broad term that can apply to research on anything connecting a concept to some unknown piece of information. Examples of this follow.

Research on referring expressions was one area in which mutual knowledge was somewhat redefined following the work of Grice. Mainly, it seems that implementations have focused on natural language generation of referring expressions (Dale & Reiter, 1995; Appelt & Kronfield, 1987; Dale 1992; Viethen & Dale, 2006). These researchers focused on computationally constructing noun phrases in ways in which a human would. Solutions to this issue varied wildly. One solution described involves defining an object as completely as possible and later removing any redundancies and over-specifications that are not necessary or understanding (Reiter, 1991). Others focused on computational complexity of Grice's Maxims, narrowing scope within referring expressions and determining what a speaker's purpose was in identifying a particular object (Dale, 1995). Though this was only related marginally to the work in this paper, it was about as close as it got to defaults. Surprisingly, these authors were not focused on the detection of mutual information/defaults but on the automatic generation of descriptions and objects that would be used to further describe the defaults.

Referring expressions was not the only area in which mutual knowledge had been implemented. Some research has been done in the area of databases to represent knowledge and use it to create richer queries. For instance, Feldman and Hirsh (1997) created a system that examines keyword labels in text documents. The system viewed background knowledge as constraints to a query.

Some research was also found in identifying the meaning of unknown words (Zhang, Zhai & Zong, 2013; Soe, 2013). One example was by Fan, Chen and Hu (2010). The basic premise behind the work of Fan et al. is that occurrences of an unknown word were compared with known words in an ontology and dictionary based on the properties of the unknown word that had been pulled from context clues. Though this was not directly related to mutual understanding it was a step towards connecting context with the properties of words and concepts.

The work of Maree and Belkhatir (2013) also showed a trend towards connecting properties with concepts. Maree and Belkhatir noted that most research on knowledge acquisition, in ontologies, is done by looking at automatic concept acquisition, finding new instances of concepts, input requirements, learning methods and the output of the ontology. Their paper differed from this trend by examining the relationships between concepts. They deemed these relationships to be missing background information not for people but for the ontology itself. Maree and Belkhatir used a combination of semantic relatedness functions and pattern acquisition to gather miss background knowledge. This was significant to this research as it showed a shift in ontology research towards pulling missing information using a corpus. In fact, it related to Ontological Semantics because properties, which represent the links between concepts, are pivotal in text meaning representation.

Loosely related to the work in this thesis was that of Balahur, Hermida, and Montoyo (2012) on implying emotion in text. Balahur, Hermida and Montoyo claimed that most sentiment is not expressed explicitly but rather implicitly. To pull sentiment, the authors created a knowledge base in the form of an ontology. The ontology was created using machine learning techniques and a large corpus called ISEAR. A portion of the data pulled from ISEAR was used to train the machine learning algorithms to create the ontology. The remaining data pulled from ISEAR was used to test the algorithms. The authors found that, even with a small data set, using an ontology in the creation and population of a knowledge base was just as successful as using supervised learning techniques. This was significant as it provided evidence and precedence for the use of ontologies in developing and storing background knowledge.

## 2.5 Ontological Semantics and Ontological Semantics Technology (OST)

One can see from section 2.4 that the research on mutual knowledge implementation has a very wide scope with very little computational research. In fact, the only approach to default identification before this investigation was in the paper by Raskin, Taylor and Hempelmann (2010). Connected with the research in this thesis, an additional two papers have discussed automatic default detection for OST (Ringenberg, Taylor, Springer & Raskin 2015; Ringenberg, Stuart, Taylor & Raskin 2015).

It was the belief of the researcher that Ontological Semantic defaults would lend themselves well to mutual Knowledge Representation because they have a strong founding in semantics and logic. As such, the researcher chose Ontological Semantics and OST as the tools for this study.

### 2.5.1 Ontological Semantics

Ontological Semantics is “a theory and methodology for representing natural language meaning” (Taylor, Raskin, Hempelmann & Attardo, 2010, p. 3335). Ontological Semantics seeks to represent the world using properties and concepts in such a way that it models human understanding. These concepts and properties are used to link sentences to their meaning in Text Meaning Representations (TMRs). Ontological Semantics is language independent; the Ontology can be used to represent multiple languages simultaneously (Taylor, Raskin & Hempelmann 2011).

Concepts within Ontological Semantics are used to represent single entities and events within the world. Concepts are part of an Ontology which links the concepts by different properties (Taylor & Raskin, 2012).

In addition to the Ontology, Ontological Semantics includes an Onamasticon, Lexicon and InfoBase. An Onamasticon is a collection of proper nouns and the properties that define them. A Lexicon is a dictionary-like collection of terms which fall under concepts in the Ontology. An InfoBase is a collection of TMRs that represent a particular dataset and the contextual connections within that dataset.

Ontological Semantics Technology is an implementation of Ontological Semantics. While Ontological Semantics is loosely based on metaphysics, Ontological Semantics Technology takes the principles of Ontological Semantics and implements them in different domains (Taylor, Raskin, Hempelmann & Attardo, 2010).

### 2.5.2 Ontological Semantics and Defaults

Defaults as described by Taylor et al. in 2010 refer to that information which is assumed to be known and is no longer salient to the speaker. Because this information is no longer salient to the speaker, it is not brought up in conversation. However, that unspoken piece of information is necessary for understanding the meaning of a statement.

Defaults marginally relate to modern research on mutual knowledge and referring expressions. Defaults are foundationally consistent with the work of Grice, Sperber and Wilson (2013, 1994).  Defaults focus on what is implied within an event. Grice was primarily concerned with the overall exchange of information and not what was mentioned or not. Referring expressions have been focused on the information provided to identify objects, not to identify what isn't described.

According to Taylor et al. (2010, p. 3334) "information in text is either just additional (previously unknown) information, or it overwrites the existing (salient) information" that has been represented by the background knowledge of the individual. The author went on to describe a means by which personal defaults should be identified using text. The authors' view was that personal defaults could be identified by looking at the values assigned to properties. Those values which are stated in text were thought to be highly unlikely to be background knowledge. If the values were considered to be background information they would not be stated. Looking at the "white dude" example again, "dude" would have been a default for the speaker. It was clear from the speaker's statement that was generally approached by males. How did we know this? If being approached by a male was novel information, the speaker would have said that a dude had approached her; the statement that the dude was white would not have been as salient (Taylor, 2010).

In 2015, further work was done on Ontological Semantic defaults (Ringenberg, Taylor, Springer & Raskin; Ringenberg, Stuart, Taylor & Raskin). These works described a partial implementation of default detection. The methods from both papers were used in

Phase 1 of this research and involved the identification of direct and indirect objects that could have been potential defaults based on the presence of an adjectival modifier.

Ringenberg, Stuart, Taylor and Raskin also identified the kinds of defaults that can occur (2015). The simplest default is the default filler which is referred to in 2004 by Nirenburg and Raskin. This is the default filler which is used to map defaults to a given property of a concept. The remaining default types include: direct object defaults, which are described in this thesis; script defaults which are loosely related to scripts as they are defined by Schank and Abelson in 1977; and semantic ellipsis defaults which are based on the work of Baltes in 1995 and McShane in 2005.

Ontological Semantics was the chosen tool of the researcher for many reasons. Ontological Semantics provided a rich platform for the relations of semantic meanings of words. As mutual knowledge focuses primarily on pragmatics, it is necessary to model not just individual words but complete sentences and ideas. Ontological Semantics was also one of the few fields in which computational research and algorithms had been outlined on this topic. As such, the researcher saw Ontological Semantics and population defaults as an excellent medium for this study.

## 2.6 Chapter Summary

In Chapter 2 the researcher discusses relevant literature to conversational inference, Ontological Semantics and defaults. Both computational and theoretical papers are discussed. In Chapter 3 the methodology for this research is discussed in detail.

# CHAPTER 3. METHODOLOGY

## 3.1 Introduction

Chapter 3 presents the methodology the researcher used to detect defaults in text. The framework, research bias, apparatus, tool creation, measurements of success, threats and weaknesses are discussed.

The methodology associated with this research was based on the algorithm for detecting WD-Inference described by Taylor, Raskin, Hempelmann & Attardo in 2010. In this paper, the authors describe a method of identifying a default by looking at the arguments to verb events. The observation is made that individuals typically only mention defaults when the default is being overwritten or novel information is being supplied. This research sought to implement, adapt and further this methodology for both direct object arguments and indirect object arguments to verb events.

The original implementation of the WD-Inference algorithm, for this research, is described in detail by Ringenberg, Taylor, Springer and Raskin with preliminary results given (2015). The implementation described in the paper nearly directly matches that of the method proposed by Taylor, et al. in 2010. The algorithm was refined and tweaked again in 2015 in the work of Ringenberg, Stuart, Taylor & Raskin. In this paper, direct object defaults were further defined within the spectrum of default types as being “conceptual objects that are inherently and obligatorily included in some verbs not (or not always) mentioned" (Ringenberg, et al., 2015). The method of implementation in this paper was very similar to the method described in the previous paper and, like the previous paper, was completed in conjunction with this research.

The methodology in Chapter 3 represents the culmination of the research from 2010 and 2015. The methods described in the first 3 papers were implemented in Phase 1 of this research. The preliminary results from Phase 1 were used to adjust the algorithm and methodology for use in Phase 2 as well as to create a tool for automatically processing defaults.

## 3.2 Framework

This thesis was a qualitative study on Ontological Semantic defaults in direct and indirect objects arguments to verb events. A qualitative study was chosen for this thesis because so little was known about default detection or the viability of automatic default detection. To the knowledge of the researcher, no previous implementations of an automatic default detection algorithm or tool had been done. As such, a narrative inquiry was suited to this investigation.

The goal of this research was to extract candidate direct and indirect object defaults from unstructured text by examining the relationships between events and the objects and fillers that describe them. Specifically, this research examined events that occured as verbs and the modifier-noun and noun combinations that were associated with them within a verb phrase. These events, modifiers and nouns were used to identify when information about an event was both stated and omitted. The purpose for stating and omitting the information, for instance in lying, was outside of the scope of this research.

This research focused on what was omitted and stated and how that information could be identified.

This research was broken into two phases. The first phase was the construction of an algorithm for identifying defaults based on verb phrases found in the Brown Corpus.

The specific verbs chosen and the quantity of verbs chosen are discussed in detail in Section 4.1. However, the goal of identifying verbs for use in this research was to create a means for comparing different implementations of the algorithms as well as to compare results from different corpora. As such, a large selection of relevant verbs was necessary.

As stated previously, this phase most closely mapped to the methodologies described in the papers available within Ontological Semantics on Ontological Semantic defaults extraction methods. The focus in Phase 1 was primarily the identification of direct and indirect objects defaults based on the presence of adjectival modifiers only. The second phase was a reworking of the algorithms in Phase 1 based upon observations made from the preliminary results. Most notably, Phase 2 introduced additional modifiers to direct and indirect objects. This change is further discussed in Chapter 4. In Phase 2, the new algorithms were run on both the original Brown Corpus dataset and also the Wikipedia data set.

It was the hope that this research would show that semantics could be used to pull defaults from unstructured text.

## 3.3 Tool Creation

For the purpose of collecting Ontological Semantic default candidates automatically, a tool was needed that could be plugged into Ontological Semantics Technology (OST) in the future. In Phase 1, all candidate defaults were pulled using computers. However, the implementation was extremely specific and stream-lined for Brown Corpus only. As a portion of Phase 2, a tool was created using the algorithms that are described in this research as well as based on the feedback from Phase 1. This tool was meant to be a general-purpose default detection tool for any corpus. Some of the key requirements for this tool included the ability to:

- Generate typed dependencies using some form of Stanford Parser and specifically collapsed-dependencies.
- Enter any number of textual documents for processing.
- Enter any number of verbs to look for within the entered text.
- Find candidate defaults using the methods described in this thesis.
- Generate output that may be analyzed outside of the tool.

The implementation of the tool is briefly described in Chapter 4.

## 3.4 Research Bias

This research was conducted by a researcher in the area of Ontological Semantics. As such all design decisions were made from the semantic perspective of the researcher. The researcher was a large part of the experiment itself. To minimize bias additional researchers within Ontological Semantics were used to examine preliminary results.

Suggestions, observations and feedback made by additional Ontological Semantics researchers were used to drive the modification of the algorithms from Phase 1 in Phase 2.

### 3.5 Apparatus

Brown Corpus and Wikipedia for Schools were the corpora used in this research. Both resources are freely available online in their entirety.

Brown Corpus was chosen because of the inherently structured nature of the documents it contains as well as due to its popularity within Natural Language Processing. Brown Corpus is a collection of documents from 1961 that have been specifically chosen for their value in comparative studies. The corpus consists "500 samples of 2000+ words each" (Francis 1979). The number of words in Brown Corpus totals 1,014,312. Documents containing over 50% dialogue were not included in the corpus. Verse was also not included in the corpus because they potentially create linguistic difficulties for researchers.

Wikipedia for Schools was chosen because of the large amount of text available. Wikipedia for Schools is a collection of over 6,000 Wikipedia documents and more than 26 million words, which all pertain to subjects taught in UK curriculum (“Wikipedia for Schools”). In addition to Wikipedia articles, additional text documents describing the hosting organization’s charities in different countries. The total document count is 8158 text document. Over 50,000 images are also included within Wikipedia for Schools.

In order to identify the target verbs in Brown Corpus, the Natural Language Toolkit (NLTK) version 2.0.4 was used. NLTK contains the tagged Brown Corpus along with the Porter Stemmer, which was used to find verb stems. Porter Stemmer was chosen because it is one of the most commonly used stemmers. Porter Stemmer was also chosen because it removed 60 endings which was far less than other stemmers. This turned out to be an advantage because it did not over-stem as much as other stemmers. Porter Stemmer is further described in the work of both Krovetz and Porter (2000, 1980). All programs the researcher wrote with NLTK used Python version 2.7. NLTK was chosen for this research because of its wide-usage within Natural Language Processing. NLTK was also used due to the vast quantity of tools available.

Stanford Parser was used in order to create dependency grammars from chosen sentences with the verbs being examined here. Stanford Parser was chosen because it provides a simple interface with which to produce dependency grammars. The researcher was aware of other dependency parsers, include Malt Parser. However, the researcher found the vast documentation and simple user interface associated with Stanford Parser to be beneficial.

Additionally, in Phase 2, Stanford's CoreNLP version 3.5.0 tool was used. CoreNLP still uses the Stanford Dependency Parser but outputs the results in a format that works more fluidly with the default detection tool. CoreNLP also allows user to take advantage of multi-threading. Stanford's website suggested giving each thread 1800MB to ensure that sentences are able to be processed. The researcher ensured that 4000MB were available to each thread to reduce the potential for memory errors.

## 3.6 Testing Methodology and Measurements of Success

Default detection is an emerging portion of Ontological Semantics. As such, very little research exists on the topic. To the researcher's knowledge, no previous algorithms had been developed for semantic default detection within Ontological Semantics before this work.

Due to the lack of previous work in this area, the measure of success for this particular algorithm was binary. The researcher only determined whether or not the algorithm was capable of detecting defaults. The emphasis of this algorithm was not on efficiency but rather accuracy. Researchers in Ontological Semantics were used to confirm, and provide feedback on, the algorithms and preliminary results from Phase 1. One of the primary measures of success was the implementation of changes to the algorithm the researchers suggested in Phase 1.

The validation of this algorithm was purposefully simple as it was an initial investigation of the phenomena of defaults. As such, the researcher needed only know that the algorithm was able to, at a minimum, make inferences about the defaults of a population. The emphasis in this research was placed on improving the methods for default detection and determining whether or not it would be viable for pulling candidate defaults that make sense to a human. Any novel information about a population's defaults that the research could generate would be useful to future researchers in this area.

## 3.7 Threats and Weaknesses

The biggest threat to this research was that it was a new approach to a largely unexplored problem. The researcher mitigated this threat by thoroughly documenting all stages of research. In this way, the researcher ensured that the research could be replicated in the future.

Another risk to this research was the availability of sufficient data. Gathering defaults for an entire population required a significant amount of data. Though semantic research had been done on background knowledge acquisition for ontologies, little research had been done on the acquisition of defaults. As such it was unclear how large a dataset was required in order to pull accurate and useful inferences. This risk was mitigated by using large, pre-existing datasets including Brown Corpus and the Wikipedia selection from Wikipedia for Schools created by SOS Children.

## 3.8 Chapter Summary

In Chapter 3, the researcher describes the methodology that will be used to detect Ontological Semantics defaults within text. The apparatus, tool and biases are all discussed.

In Chapter 4, the researcher discusses the results coming from both Phases 1 and 2 in detail.

# CHAPTER 4. DATA ANALYSIS

This chapter presents the findings for different methods of identifying candidate defaults within text. Additionally, details regarding the changes made between Phase 1 and Phase 2 are discussed. The differences in Phase 1 and Phase 2 are ultimately compared.

The overall approach to the detection of defaults was to analyze raw data, taken from a popular and well cited corpus, in an attempt to generalize the thought process used by human beings to understand unspoken information. These generalizations were used to develop an algorithm capable of identifying basic defaults in most genres of text. The focus of Phase 1 was on the detection of adjectival modifiers of direct and indirect objects of an event. The focus of Phase 2 was on the detection of several modifiers of direct and indirect objects. The focus in Phase 2 shifted from Phase 1 due to the data analysis of Phase 1. Further details on these changes are presented below.

## 4.1 Process: Phase 1

The first phase of the research was used to both generate a set of data for testing and to create basic, generalized rules about the relationships between verbs and adjective-noun and noun combinations that relate to defaults.

### 4.1.1 Verb Selection within Brown Corpus

In order to find defaults, a set of verbs was needed to use consistently throughout the research. For this purpose, Brown Corpus was used. Brown Corpus was chosen because it was tagged for part-of-speech (POS) by humans.

Initially, all tokens tagged within Brown Corpus as verbs were pulled. Each verb was stemmed using Porter Stemmer. A stemmer was used to attempt to combine verbs in different forms. If this step had not been added, “drive” and “driving” would have been seen as two independent verbs.

Once all verbs were stemmed, the frequency of each of the unique stems was calculated. The top two hundred most frequent verbs were chosen from this list. The researcher chose to pull only the 200 most frequent verbs because the frequency with which stems occurred within Brown Corpus significantly declined after the 200th stem.

Entries in the 200-verb-list were removed if they were believed to be irrelevant to the discussion. Verbs such as “to say”, “to do”, “to go” and “to be” were removed as they traditionally appear in most stop lists within Natural Language Processing algorithms. Verbs were also removed from consideration if the stemmer stemmed them down to 1-2 letters. The remaining verbs were then reviewed by the researcher and any duplications that remained, due to verb forms not being properly stemmed, were combined. The final list of verbs to be used for default analysis totaled 145 verbs.

All sentences within the Brown Corpus that contained these verbs as verbs were used. This means that sentences that contained the verb “walk” were pulled but sentences containing “walk” as a noun were omitted.

### 4.1.2 Using Stanford Parser to Pull Dependencies and Tags

Stanford Parser was used on the sentences pulled from Brown Corpus to create dependencies for each sentence. The *nsubj, dobj, iobj* and *amod* tags were used to pull the relevant data. The *nsubj, dobj* and *iobj* tags were used to pull all verbs within a sentence. The *dobj* and *iobj* tags were used to pull direct object and indirect object nouns associated with the verb as well. The *amod* tag was used to connect an adjective with a noun.

The procedure for pulling the appropriate information from a typed dependency consisted of the following steps:

1. Pull all *nsubj* tags from a given dependency**.** If the *nsubj* tag contains one of the verbs in the investigation, store the verb in the *nsubj* tag. If the *nsubj* tag doesn't contain an appropriate verb ignore the *nsubj* tag.
2. Pull all *dobj* and *iobj* tags from a given dependency. If the first argument of the tag is one of the target verbs, store both arguments. This includes the verb and the direct or indirect object.
3. Pull all *amod* tags from a given dependency. Store all *amod* tags for the dependency. The *amod* tag includes a noun and the adjective that modifies the noun.
4. Compare the list of *dobj*/*iobj* tags with the list of *amod* tags. Link all instances where the noun indexes are the same in both tags and the noun in both tags is the same. If a noun in a *dobj* or *iobj* tag has no adjective modifier, save it to a separate list.
5. Compare the list of links from the previous step to the list of *nsubj* verbs. If the ID and verb in the *nsubj* tag, the first argument, matches that of the

verb in the *iobj* or *dobj* tag, remove the verbs from the list of *nsubj* verbs. This ensures that the count of verbs that occurred with no arguments is accurate.

The output of this process was 3 separate lists: verbs with no arguments, verbs with only a noun argument and verbs with a noun argument which was linked to an adjectival modifier.

Total, there were 13,493 unique instances of verbs being used with no arguments at all. A total of 8,190 verb instances with only a noun argument were found and 2,556 verb instances with a noun and an adjective were found.

### 4.1.3 Comparing Verbs, Verb-Nouns and Verb-Noun-Adjectives for default candidacy

The lists in the previous section were compared. Consistent with the methods described in the 2010 and 2015 papers related to defaults, verb-noun combinations that existed in both the verb-noun list and the verb-noun-adjective list were removed as potential defaults (Taylor, Raskin, Hempelmann & Attardo; Ringenberg, Stuart, Taylor & Raskin; Ringenberg, Taylor, Springer & Raskin). These entries were removed because this demonstrated that the noun was being used without modifiers. Nouns that occur without modification are unlikely to be defaults. For instance, one would not find oneself saying “I eat food” generally. This is because “food” is implied within the event-concept of “eat” (Ringenberg, Stuart, Taylor & Raskin, 2015). Verb-noun combinations within the verb-noun-adjective triples that did not exist within the verb-noun list were flagged as candidates. The data pulled from this phase was sorted by verb and then by noun.

Total, 2,234 instances of potential candidates were found with this method. This is across the approximately 20,000 sentences that were originally pulled from Brown Corpus with the target verbs. A summary of the data follows in Table 4.1.

Table 4.1. *Summary of Phase 1 Results*

| Metric | Count |
|---|---:|
| Total Verbs Chosen | 145 |
| Total Documents Analyzed | 500 |
| Verbs with No Modifiers | 13493 |
| Verbs with Noun Modifier | 8190 |
| Verbs with Noun and Adjective Modifiers | 2556 |
| Total Instances of candidates | 2234 |
| Unique Verb Forms with Instances of Candidates | 449 |

In this analysis, verbs were not aggregated by verb infinitive. As a result, it was difficult to determine the total number of unique verbs that were found to have candidate defaults. Upon first glance, there were 449 unique verb forms with candidate defaults. However, this means that “given” and “gave” were considered to be separate verbs. Even so, several verb form were found to have multiple potential defaults. Of all the unique verb forms, 201 verbs had only one instance of the verb having a potential default. This left 246 verb forms with multiple instances of candidate defaults. Within each verb, one could see both entirely unique noun defaults and multiple instances of the same default. The top 20 verbs with the highest number of candidate defaults are shown in Table 4.2.

Of the 20 verbs in Table 4.2, 12 had multiple instances of the same default appearing in Brown Corpus. This is unsurprising due to two factors: the size of Brown Corpus and the verbs that appeared.

A significantly larger corpus was chosen in Phase 2 to confirm that automatic direct and indirect object default detection was possible. This is because the researcher suspected that frequency of defaults would potentially be low in Brown Corpus.

Table 4.2 *20 Verb Events with Highest Number of Candidates (Brown Corpus – Phase 1)*

| Verb | Count of Candidates |
|---|---:|
| Left | 20 |
| Need | 21 |
| Held | 21 |
| prevent | 21 |
| Seen | 21 |
| build | 22 |
| keep | 23 |
| maintain | 23 |
| bring | 25 |
| offer | 25 |
| Felt | 26 |
| develop | 27 |
| brought | 28 |
| Saw | 36 |
| given | 52 |
| Use | 52 |
| See | 59 |
| gave | 87 |
| made | 105 |
| make | 121 |

Also, the verbs that had the largest number of instances of candidate defaults were verbs that tend to have many different lexical senses in general. Presumably, each lexical sense should have had at least a single default if not multiple. Verbs like “to make” and “to give” have extremely wide scope and therefore showed up with larger amounts of candidate defaults than other verbs. This was expected and further validated the decision to pull such a large number of verb events.

### 4.1.4 Researcher Feedback

In order to confirm and alter the algorithms in Phase 1, in accordance with the Measures of Success discussed in Chapter 3, researchers within Ontological Semantics were assembled. Initially 6 researchers, including the author discuss the results. Later, the researcher and 3 other researchers gather the results. Partial and full results of Phase 1 are noted in 2015 (Ringenberg, Stuart, Taylor & Raskin; Ringenberg, Taylor, Springer & Raskin).

The observations made by the researchers include the following:

- The algorithms did pull defaults that were based on adjectival modifiers. However, potentially more than half of all direct and indirect object defaults were being omitted by not including additional modifiers.
- Events should be aggregated as candidacy for being a default is not determined by the form of the verb but by the event that the verb represents.
- candidates triples seemed to fall into a few common patterns including:
    - Events with several default candidates
    - Events with few, and unexpected, candidates
    - Events with entirely expected defaults

## 4.2 Process: Phase 2

The second phase of the research was used to test the modified algorithm created in Phase 1 and analyze the ability of the algorithm to both detect defaults in Brown

Corpus and to generalize to other corpora. In this thesis, the Wikipedia dataset was used to analyze the effectiveness of the implemented algorithms.

### 4.2.1 Changes from Phase 1

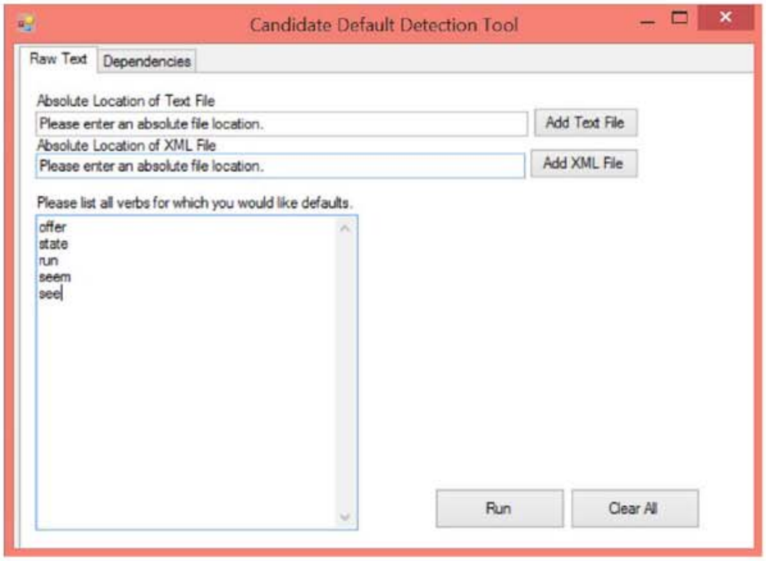
As a result of the analysis from Phase 1 the algorithm was altered slightly in this phase. Primarily, the analysis was changed from strictly adjectival modifiers to a noun to all modifiers to the noun except for determiners. A significant portion of potential candidate direct and indirect object defaults were ignored when only adjectival modifiers were used. In order to better understand direct and indirect object defaults, it was necessary to broaden the scope of modifiers. The modifiers added in this phase were: *appos, advcl, predet, preconj, vmod, mwe, advmod, rcmod, quantmod, nn, npadvmod, num, number, prep* and *possessive.*

An additional tag was also added to the algorithm to pull verbs from text: *nsubjpass.* In phase 1, the research realized that verbs were potentially being omitted by not pulling them from passive clauses.

The method by which verbs were compared to a given corpus also changed as a result of the analysis from Phase 1. In Phase 1, the algorithm did not look for all forms of a given verb in the corpus. Thus, sentences with “make” would be included in the analysis but sentences with “made” would not necessarily be pulled from the corpus. Using the tool, all forms of both regular and irregular verbs were examined and aggregated.

### 4.2.2 Tool

Additionally, a tool was developed to automate the modified process from Phase 1. The tool included the ability to create Stanford Dependency parses; pull *nsubj,* *nusbjpass*, *dobj*, *iobj* and modifier tags; link verbs, noun and modifiers in the method described in Phase 1; and identify and save candidate defaults from the data set. The tool also allowed the user to enter any desired verbs for which to find defaults. The interface for verb and text document entry is shown in Figure 4.1 below.

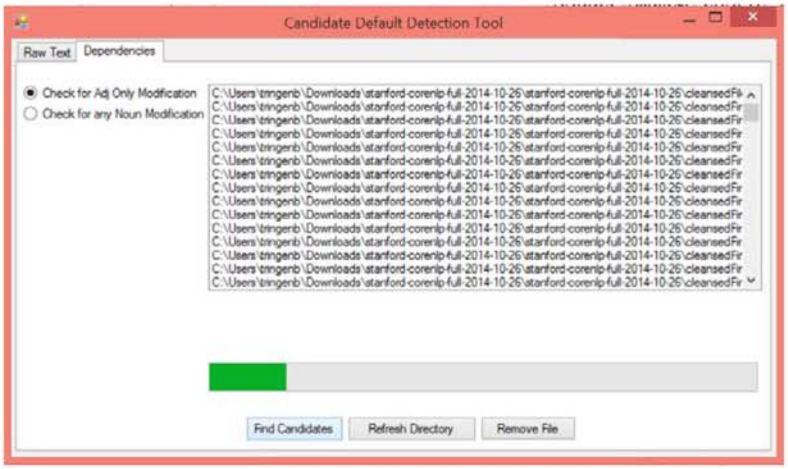


Absolute Location of Text File
Please enter an absolute file location.
Absolute Location of XML File
Please enter an absolute file location.

Please list all verbs for which you would like defaults.
offer
state
run
seem
see

*Figure 4.1* Data entry portion of the default detection tool

The ability to identify defaults using only *amod* as well as to identify them using all other modifiers, except for determiners, was included within the tool. Figure 4.2 shows the dependency analysis tab within the default detection tool.



● Check for Adj Only Modification
○ Check for any Noun Modification

```
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFil
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
C:\Users\tringenb\Downloads\stanford-corenlp-full-2014-10-26\stanford-corenlp-full-2014-10-26\cleansedFi
```

*Figure 4.2* Dependency analysis portion of the default detection tool.

### 4.2.3 Collecting and Cleaning Wikipedia Data

The tool created in Phase 2 was used to run the Wikipedia corpus. The Wikipedia for Schools corpus was chosen for this thesis because of the size and abundance of texts. More information on this corpus and why it was been chosen is available in Section 3.4.3.

For this phase, the entire Wikipedia selection was downloaded. All html documents were originally pulled from the selection. PDF documents and image files were ignored for the purposes of this thesis. A total of 8158 html documents were found.

The html documents found in the Wikipedia selection were then cleaned. Initially only the html tags were removed. However, upon running an html file, cleaned of only tags, through the Stanford Parser it became evident that further cleaning was necessary.

In addition to removing html tags, links were removed from the data when possible. This was essential to processing the documents through the Stanford Parser as the dependency parser occasionally tried to find dependencies between each character of a link. This produced unnecessary noise within the data and also significantly increased the time required to parse a single document.

Consistently occurring, non-numeric or alphabetic, characters were also removed from the data. This included the following: [, ], #. These characters were only removed in sections of the document that were consistent through all documents. For instance, section headers contained [ and ]. These characters were treated as tokens and placed within the typed dependency. Though not all instances of this could be removed, as many as possible were in order to again reduce noise.

In cleaning the data, it was found that 181 of the documents included in the corpus contained encoding issues. As a result, they were removed from the investigation. An additional 3 documents were found to be empty and contained no data. These 3 documents were also removed from the investigation bringing the total number of documents available for analysis to 7974.

### 4.2.4 Parsing and Candidate Detection Using the Tool

The documents that remained were parsed using the Stanford Dependency parser within Stanford CoreNLP. The parser that was used in Phase 2 was the same as the parser in Phase 1. CoreNLP was specifically used for Phase 2 because the dependencies were produced in XML which the researcher felt was easier to work with than the output of the Stanford Parser by itself.

Finally, the tool was used to create candidate defaults from the Wikipedia corpus. Both the adjective-only method and the all-modifiers method were run using the tool. The purpose of this was to ensure that the changes suggested by Ontological Semantics researchers were appropriately made between Phases 1 and 2. Complete results from all phases are available through the researcher. The following tables briefly summarize the data from the adjective-only method and the all-modifiers method.

Table 4.3 *Summary of Wikipedia Adjective-Only Analysis*

| Metric | Count |
|---|---:|
| Total Verbs Chosen | 145 |
| Total Documents Analyzed | 7974 |
| Verbs with No Modifiers | 308689 |
| Total Instances of candidates | 88948 |
| Unique Verbs with Candidates | 135 |

Table 4.3 shows some high-level information about the adjective-only Wikipedia analysis. As stated previously, this data was collected in a way similar to the process described in Phase 1. This method still only used adjective modifiers but aggregated the data by verb infinitive to get a clearer picture of what is happening. As the Wikipedia documents were treated as objects in the tool, the problem of pulling verbs that only contained the relevant verbs was no longer a problem. An example of this would be having a verb “run” and pulling entries for “overrun” because “overrun” contains the other verb. This was an issue in Phase 1 but was not in Phase 2.

Substantially larger numbers of candidate defaults were found in this data set over Brown Corpus. However, this was to be expected due to the sheer size of the Wikipedia data set in comparison to Brown Corpus.

It was also interesting to note that most of the verbs that were used in this analysis had at least 1 candidate default as 135 unique verbs had candidate defaults associated with them.

The following table breaks down the same information from Table 4.3 for all modifiers instead of just adjectives. It is important to reiterate here that adjective modifiers were still included in this analysis but were not the only factors in determining default candidacy any longer.

Table 4.4 *Summary of Wikipedia All-Modifiers Analysis*

| Metric | Count |
|---|---:|
| Total Verbs Chosen | 141 |
| Total Documents Analyzed | 7974 |
| Verbs with No Modifiers | 272216 |
| Total Instances of candidates | 205774 |
| Unique Verbs with Candidates | 141 |

As can be seen in Table 4.4, the number of documents analyzed remains the same and yet the total number of candidate defaults had more than doubled. If the adjective-only methods were to be removed from this analysis, 116,826 candidate defaults would remain as a result of the remaining modifier tags.

The 20 verbs with the highest occurrence of default candidate instances are shown in the tables below for both the adjective-only method and the all-modifiers method.

All 20 verbs in Table 4.5 had multiple instances of the same meaningful candidate defaults appearing in Brown Corpus. Meaningful was defined in this context as being the result of a logical dependency and not on an error of the dependency parser. For instance, the verb “use” had 157 instances of the character _ being a candidate default. Upon reviewing the data, this appeared to be the result of the parser attempting to parse any remaining hyperlinks as well as parsing list characters.

The abundance of candidates and the higher frequencies for each candidate was expected due to the size of the corpus.

Table 4.5 *20 Verb Events with Highest Number of Candidates (Wikipedia – Adjective Only Method)*

| Verb Event | Count of Candidate Default Instances | Count of Unique Candidates |
|---|---|---|
| Achieve | 1262 | 434 |
| Describe | 1286 | 690 |
| Reach | 1660 | 539 |
| maintain | 1699 | 637 |
| support | 1887 | 741 |
| require | 1890 | 859 |
| Allow | 2100 | 1053 |
| establish | 2126 | 663 |
| Offer | 2342 | 712 |
| consider | 2370 | 1075 |
| develop | 2629 | 900 |
| receive | 3044 | 836 |
| Play | 3079 | 415 |
| Cause | 3148 | 1021 |
| contain | 3639 | 1359 |
| Form | 3905 | 1254 |
| Create | 3963 | 1551 |
| produce | 4231 | 1563 |
| include | 5853 | 2678 |
| Use | 8742 | 2718 |

Though many of the top 20 verbs were still verbs that tend to have several senses, there appeared to be more verbs with meaningful defaults in the top 20 verbs of the Wikipedia corpus over the Brown Corpus.

Table 4.6 shows the 20 verbs with the highest frequency of candidate defaults. All of the noun modifiers mentioned in this research were used to find the candidates in this dataset.

Several verbs in Table 4.6 are different from the verbs in Table 4.5. Adding in additional modifiers significantly affected the amount of candidates that were able to be pulled from the same corpus. Even for the verbs that were the same between the 2 methods, it was clear that looking at all modifiers produced more candidates. As an example, the verb event “play” had 3079 instances of candidates in the Wikipedia texts when only the adjectival modifier was used. With all modifiers, 5343 instances of candidates were found for “play”.

Table 4.6 *20 Verb Events with Highest Number of Candidates (Wikipedia – All Modifiers)*

| Verb Event | Count of Candidate Default Instances | Count of Unique Candidates |
|---|---:|---:|
| Include | 21473 | 5740 |
| Use | 20187 | 4178 |
| produce | 7869 | 2112 |
| form | 7763 | 1708 |
| contain | 7662 | 2034 |
| create | 7510 | 2039 |
| receive | 6252 | 1163 |
| cause | 5746 | 1566 |
| play | 5343 | 843 |
| develop | 5044 | 1165 |
| establish | 4738 | 989 |
| allow | 4659 | 1779 |
| support | 4620 | 1182 |
| reach | 4609 | 954 |
| consider | 4013 | 1458 |
| offer | 3896 | 960 |
| require | 3253 | 1210 |
| maintain | 3087 | 830 |
| describe | 2981 | 1131 |
| Kill | 2713 | 705 |

For consistency, the sentences pulled from Brown Corpus were also re-examined using the adjective-only and all-modifier methods. Summaries of both sets of data are presented in the tables below.

Table 4.7 *Summary of Brown Corpus Adjective Only Modifiers Analysis*

| Metric | Count |
|---|---:|
| Total Verbs Chosen | 141 |
| Total Documents Analyzed | 500 |
| Verbs with No Modifiers | 1488 |
| Total Instances of candidates | 398 |
| Unique Verbs with Candidates | 107 |

The default detection process that is represented in Table 4.7 is the same as that in Table 4.1. The only differences are those changes which were suggested in Phase 1. The reason for the changes in the number of candidates per verb was that the verbs in this section were aggregated. The default detection tool also ensured that only verbs that were chosen for this analysis were truly used as well. Thus, the data represented the data from Phase 1 but was slightly less noisy. However, the core of the data and methodology remained unchanged.

The table below shows the top 20 verbs with the most instances of candidate defaults in Brown Corpus when the adjective-only method was used. The data appeared to be different from the data in Table 4.2. Again, this was only due to aggregation. Table 4.2 shows verb forms while Table 4.8 shows the verb forms aggregated into the verb of which the verb form is a part.

Table 4.8 20 *Verb Events with Highest Number of Candidates (Brown – Adjective Only)*

| Verb | Count of Candidate Default Instances | Count of Unique Candidates |
|---|---:|---:|
| use | 29 | 26 |
| develop | 24 | 17 |
| add | 16 | 11 |
| offer | 14 | 13 |
| need | 14 | 13 |
| form | 12 | 8 |
| enjoy | 12 | 9 |
| play | 12 | 12 |
| present | 11 | 9 |
| face | 11 | 9 |
| allow | 10 | 9 |
| accept | 9 | 9 |
| carry | 8 | 7 |
| produce | 8 | 7 |
| note | 8 | 6 |
| include | 7 | 6 |
| pick | 7 | 6 |
| remove | 7 | 6 |
| prevent | 6 | 6 |
| expect | 6 | 5 |

Table 4.9 includes a basic summary of the information from using the all modifiers algorithm on Brown Corpus. As in all previous data the same 141 verbs were used for pulling the defaults.

As is evident in the analysis of the Wikipedia corpus, pulling all modifiers for Brown Corpus produced much higher quantities of candidate defaults in general.

Table 4.9 *Summary of Brown Corpus All-Modifiers Analysis*

| **Metric** | **Count** |
|---|---:|
| Total Verbs Chosen | 141 |
| Total Documents Analyzed | 500 |
| Verbs with No Modifiers | 1339 |
| Total Instances of candidates | 868 |
| Unique Verbs with Candidates | 106 |

The top 20 verbs with the most instances of candidate defaults within the Brown Corpus data using the all modifiers method are shown in Table 4.10. This table shows that frequencies of candidate defaults per verb, when the all modifiers method was used, were much higher in Brown Corpus as well.

Table 4.10. *Verb Events with Highest Number of Candidates (Brown Corpus – All Modifiers)*

| Verb | Count of Candidate Default Instances | Count of Unique Candidates |
|---|---:|---:|
| use | 78 | 50 |
| develop | 41 | 27 |
| need | 36 | 25 |
| add | 29 | 16 |
| offer | 27 | 21 |
| play | 27 | 17 |
| enjoy | 22 | 13 |
| form | 20 | 10 |
| face | 19 | 13 |
| cover | 18 | 15 |
| present | 16 | 12 |
| end | 16 | 11 |
| include | 15 | 8 |
| remove | 15 | 8 |
| reach | 15 | 8 |
| watch | 15 | 6 |
| accept | 14 | 12 |
| pick | 14 | 9 |
| start | 14 | 9 |
| produce | 13 | 8 |

## 4.3 Chapter Summary

In this chapter, the researcher details the procedures used in identifying candidate defaults for each verb chosen for this investigation. Results for each method are presented and changes made between Phase 1 and Phase 2 are discussed.

# CHAPTER 5. DISCUSSION, CONCLUSIONS AND FUTURE WORK

This chapter summarizes the findings and the analysis within this thesis. Future work in the area of default detection is also briefly discussed.

## 5.1 Discussion and Conclusions

In this section the researcher addresses each of the points, in the researchers' feedback, in Chapter 4:

- The algorithms did pull defaults that were based on adjectival modifiers. However, potentially more than half of all direct and indirect object defaults were being omitted by not including additional modifiers.

This phenomena was seen in both the adjective-only data for Brown Corpus and Wikipedia but was especially true for Brown. Originally, the researcher was unaware of how few lexical entries fell within adjectival modifiers in Stanford Parser. This feedback was used to pull additional modifiers in Phase 2.

Below is an example, from the Wikipedia corpus, of when a different modifier was required in order to pull the correct potential default. This was actually an interesting example because it showed that using additional modifiers to pull defaults also helps to pull incorrectly labeled dependencies.

"By working with SOS Children as your charity partner you can place Corporate Responsibility at the heart of your business, like these companies have done …"

For this particular sentence, “responsibility” was pulled as a potential default for the verb “place”. This made sense to an English speaker as “place” and “responsibility” are closely linked in one of the senses of “place”. However, Stanford Parser labeled “Corporate*”* as a noun. So, the modifier used to pull this relationship was *nn*. So, this instance of “responsibility” would not have been recorded as a default for “place” even though a speaker would recognize it as such.

The researcher also believed that “Corporate” should have been labeled by the parser as an adjective. However, no parser is perfect. Though Stanford Parser did not always get the correct dependency, it was very good at determining that there was a dependency. So, pulling all modifiers helped to remove false negatives.

- Events should be aggregated as candidacy for being a default is not determined by the form of the verb but by the event that the verb represents.

This mistake was handled between Phases 1 and 2. The candidates were all aggregated by infinitive as opposed to verb form. This significantly helped to clarify the results. The new method, though a seemingly small change, was better suited towards default detection.

The problem of pulling non-selected verbs was also addressed in Phase 2. The method by which the information was pulled was different in the tool. An XML object was created from the document that contained different tags representing the different relationships and tokens. As a result, the researcher was able to pull only those events that exactly matched each form of the chosen verbs. This removed the noise. In the first analysis, a lot of the potential candidates were, indeed, candidates but were not for the verbs chosen. The researcher believed that this was partially due to an error in the original code and partially due to smaller verbs being contained in longer verbs. For instance, the verb “give” would not only pull the sentences with “give” in it but also “forgive”. Again, this problem was handled in Phase 2.

One problem that was unable to be addressed in this work was the issue of plural and singular nouns. Originally, the researcher attempted to remove the “s” ending from all nouns. This was not entirely successful. The next step was removing “es” and “s” from the nouns. Again, this resulted in some positive data and some rather horrible data. Finally, Porter Stemmer was used. This, also was unacceptable as many of the nouns were truncated to the point of not being distinguishable. This ruled out the use of a stemmer as Porter Stemmer removes far fewer endings than most other stemmers; the researcher ultimately felt that leaving singular and plural nouns separate was more beneficial to the overall analysis.

- candidates triples seemed to fall into a few common patterns including:
    - events with several default candidates
    - events with few, and unexpected, candidates
    - events with entirely expected defaults

There were many potential reasons for the existence of events with several candidate defaults.

One reason was that defaults differ slightly from person to person. As stated previously, most people think of “food” being the implied object for “to eat”. Many people also think of “car” being the implied instrument of “to drive”. However, that doesn’t mean that this is consistently the case. Some people may consider driving motorcycles or boats to be more obvious. The researcher expected that instances such as these would have a very low frequency occurrence in comparison to more likely candidates. In fact, that was what the data showed. Candidates that the general population would consider to be the most obvious seemed to have high frequencies of occurrence.

Take the example of “to play”. What comes to mind with the verb event “to play”? For the research, games, sports, instruments, acting and music came to mind. Table 5.1 shows the most frequent noun candidates for “to play” in the Wikipedia all-modifiers dataset.

Many of the defaults, chosen above, for the verb event “to play” occurred in the most frequent candidates above. “Role”, “roles”, and “part” all link to the sense of “play” that involves acting. “Games”, “game”, “match”, “cricket”, “football”, “tournament” and “ball” all fall under the concept of GAME. “Instrument”, “notes” and “music” all fall under the concept of MUSIC. “Members” are potentially agents involved in the concept of PLAY. “Victory” is one of a fuzzy set of outcomes for the concept of PLAY.

This was less evident within the dataset for the all-modifiers version of the Brown results. However, the researcher believed that this was again because of the size of the corpus. Interestingly, the most frequent noun candidates for “play” when using the all-modifiers method on Brown Corpus included: “swing” (6 occurrences), “golf” (2 occurrences), “jazz”(2 occurrences), “course” (2 occurrences), and “cards” (3 occurrences). Though these are not exactly what a human sees as implied in the verb “play”, all of these nouns map to an ontological concept that is mapped as a default to "play". This was expected as lexical items have different senses. This observation is discussed in detail below.

Table 5.1 *Most Frequent Candidate Defaults for Play*

| Noun | Count of Candidate Default Instances |
|---|---:|
| Role | 1645 |
| Games | 245 |
| Part | 235 |
| Game | 168 |
| Roles | 149 |
| Match | 144 |
| Cricket | 51 |
| Music | 44 |
| Victory | 38 |
| Tournament | 36 |
| Football | 34 |
| Members | 34 |
| Instrument | 32 |
| Notes | 29 |
| Ball | 28 |

The initial results from Phase 1 for Brown Corpus were difficult to interpret due to the size of the corpus. Ignoring the fact that all manner of verbs were somehow pulled we still saw high variance among potential defaults. In Phase 2 there was still a high variance in potential defaults per verb. However, the frequencies of the defaults provided insight. With the results in Brown corpus the frequencies were far lower per default. There just was not enough data within Brown to see significant enough patterns. This was why the analysis with Wikipedia was so crucial. The target sentences within Brown Corpus consisted of approximately 20,000 sentences only which amounted to a few MB of data. The full data from Wikipedia was around 8 GB.

Furthermore, when changes were made to the adjective-only algorithm in Phase 1 it resulted in more meaningful results but also resulted in about a tenth of the data. This was huge as it meant that there were even fewer defaults with even lower frequencies.

The verb event “design” was an excellent example that demonstrated the need for large datasets for default detection. The defaults for “design” were related to buildings/structures, systems and plans. Some would argue that a structure could be a complex system but for the purpose of this paper the researcher considered them to be separate categories. Even so, they both could be considered defaults for “design”.

Tables 5.2, 5.3 and 5.4 show the results from using the adjective-only method on Brown Corpus in Phase 1, Brown Corpus in Phase 2 and the Wikipedia Corpus also in Phase 2 for “design”. In Phase 1, only 2 potential candidates were found for “design”: “Buttresses” and “Scheme”.

Table 5.2 *Results for Design in Brown Corpus with Adjective-Only Method in Phase 1*

| Frequency | Nouns |
|---|---|
| 1 | Buttresses, Scheme |

In Phase 2 only a single candidate was found in Brown Corpus*:* “Buttresses”.

Table 5.3 *Results for Design in Brown Corpus with Adjective-Only Method in Phase 2*

| Frequency | Nouns |
|---|---|
| 1 | Buttresses |

These were fairly decent candidates. A buttress is a support that is built against a wall. This maps to some sort of BUILDING-PART in the ontology. A BUILDING-PART is indeed a structure.  A “scheme” is a PLAN which makes it correct in terms of what people may consider as a default for “design”. However, this required a human analysis. Would this still be definitive when done by a computer? No. There were not enough instances of either of these defaults to make them stick out as obvious. One instance of each default did not instill much confidence in them.

The results from Wikipedia were another story. Since the Wikipedia corpus was so large there were higher distributions for each candidate. As an example, “buildings” and “system” were the top 2 most frequently occurring defaults. These were the defaults a speaker would choose.  Nouns such as “language”, “scripture”, “area” and “action” all made sense as direct objects of “design" but were not defaults. Thus, the quantities for these candidates were expected to be fairly low and in fact they were. As one can see, there were few frequently occurring defaults which is what the researcher was looking for.

Another reason for the existence of so many defaults was the polysemous nature of these verbs. Taking the verb “to play” again, it was evident that several of the frequently occurring candidates made sense; “games”, “sports”, “instruments” and “roles” all showed up as frequent for “to play”. This was because these all map to different concepts within an Ontology. Playing an instrument and playing music are not necessarily the same event. That is why they are discussed separately in the explanation of Table 5.1. Playing an instrument implies that there is a person physically manipulating an instrument in such a way that it makes noise. Playing music can imply that some is physically engaging an instrument in a methodic way or it can merely mean that a person is actively listening to music which they have selected themselves. As most of the verbs in this study were highly polysemous, having multiple potential candidates with high occurrence for each verb was expected. In future work, the researcher would like to map these lexical senses and defaults to their respective ontological concept within OST.

Table 5.4 *Results for Design in Wikipedia Corpus with Adjective- Only Method*

| Frequency | Nouns |
|---|---|
| 26 | Buildings |
| 12 | System |
| 7 | Car, Church, Systems |
| 5 | Complex, Language, Scripture |
| 4 | Automobile, Building, Capital, Cars, Churches, Circuit, Computer, Ft, Garden, Helicopter, Locomotive, Mansion, Propellers, Version |
| 3 | Box, Bridge, Clock, Costume, Disc, Flyer, Plan, Structures, Swimsuit, Symbols, Vehicle, Weapons |
| 2 | Airships, Area, Arena, bank, Bareback, Bomb, Bridges, Calculator, Cathedral, Centre, Chaps, City, Cockpits, Conveniences, Engine, Exchange, Expansion, Façade, Fluoroscope, Frescoes, Gallery, Gardens, Heart, Homes, Huts, Interior, machine, Mechanism, Methodology, Museum, Network, Programme, Promenade, Pump, Research, Ring, Ship, Solutions, Statues, Submarines, Typeface, Unit, Woodcuts |
| 1 | A, Action, Agent, Aircraft, Antibodies, Apparatus, Appearance, Arches, Balloons, Basilica, BT34, BT48, Cartridge, Chapel, Class, Clothing, Coat Computers, Console, Covers, CPU, CPUs, Decoration, Edges, Equipment, Experiments, Façade, Factories, Figures, Flag, Gaol, Glass, Government, Gun, Halls, Hardware, Hooks, House, Hundreds, Hybrides, Images, Items, Itinerary, Kind, Lincoln-Zephyr, Logo, Machines, Marques, Materials, Memorial, Method, Metres, Microscope, Mimics, Mine, Model, Module, Mosaics, Motor, orangery, Order, Palace, Panther, Parts, pavilion, Pieces, Plane, Plant, Plants, Policies, Press, Process, Processor, Products, Projection, Projects, Range, Roles, Rooms, Seating, Section, Series, Shapes, Stage, Statue, Strategy, Subjects, Submarine, Supercomputer, Supply, Sybmol, T-43, Terminal, Tools, Tractor, Truck, Trumpet, Type, Urbanization, Variants, Variation, Works |

As somewhat of the inverse of the reason above, there were also so many candidate defaults because the nouns in this study were merely lexical items. Several of the lexical items for each verb truly map to the same concept. For instance, using the verb “to play” again in the Wikipedia dataset, there was a very high occurrence of the noun candidate “game”. However, “roulette”, “pokemon”, “pong”, “pac-man”, “mini-games”, "match", "hockey" and "games" all showed up as candidates for "play" as well. Many of these had high frequency of occurrence and they were all instances of the same concept GAME. Because of this there were significantly fewer candidates for each concept than there seemed to be upon first glance. Another example of a verb that fit this pattern was "express". Within Brown Corpus, the all-modifiers method pulled 3 candidate defaults: "desire", "fears" and "thanks". All 3 of these words are emotionally charged and fit under a parent, or possibly higher ancestor of "emotional-states". The results from Wikipedia showed even clearer results with candidate defaults including: "abhorrence", "admiration", "affection", "fears" and "feelings".

Finally, there was the fact that several of these candidate defaults actually map to different properties of a concept. "Car", for instance, is a potential default filler for the property of instrument for the concept of DRIVE. "Student", on the other hand, is a potential default filler for the property of beneficiary for the concept of TEACH. This observation was consistent with the data. However further inquiry is desired.

Events with few and unexpected candidates were largely seen within the adjective-only results. The verb event "to cut", specifically in the Wikipedia adjective-only data set, was a great example of this. The researcher would expect the defaults for "cut" to include things like "food", "knife" or "time". However "a", "lakes" and "miles" were the only defaults found. The only one of these that made any sense was miles because it is a unit of measurement. It seemed that this occurred with "cut" and other verbs because so much was left out when only looking at adjectival modifiers. One cannot have a good understanding of what is and isn't important when only looking at a single modifier.

As mentioned earlier, the candidates with low frequencies were likely to not be true defaults for a given event. This was another possible reason for few and unexpected candidates. There was the potential that some of the nouns with low frequency of occurrence just were not truly defaults or were defaults for a very small subset of the population. This could be seen for several of the low frequency candidates in Table 5.3.

Many of the verbs had entirely consistent defaults; “play”, “design”, “marry” and “recognize” are all 3 examples. This was evident in both Phase 1 and Phase 2. The researcher believed that the defaults pulled with the all-modifiers method, for both corpora, were the most representative of defaults in the data. However even with the adjective-only method, it was clear that pulling nouns that link to a verb and were never seen unmodified produced viable candidate defaults. With the changes that were suggested in Phase 1, it appeared that default detection was possible and showed promise.

As a note, the researcher did not and does not believe that a default has to never occur within a text. It is possible that defaults will occur as direct object, both modified and unmodified, on occasion. However, this research focused merely on identifying cases where there were no instances of the default. This was because it was unclear what the threshold for remaining a default would be if the default were mentioned in text. What ratio of stated versus unstated defaults is appropriate to still determine candidacy? It is unclear. This will require additional research in the future.

## 5.2 Future Work

The goal of this research was to create, improve and implement a method for the detection of Ontological Semantic defaults. The researcher believes that the focus of future work in the area of defaults should be on detecting other forms of defaults, linking candidate defaults to the Ontology and using the algorithms in this paper to examine default violations by individuals.

The type of default investigated in this work is just one in a set of four types of defaults that have been identified recently (Ringenberg, Stuart, Taylor & Raskin 2015). Future work should expand the implementation of default detection to the other three areas of defaults.

Now that a method has been implemented for identifying candidate population defaults, it is important that future work use these potential defaults to either populate property fillers of verb-events or use them to infer information about verb-events.

Lastly, future work should focus on using the algorithms described in this thesis for detecting default violations. The output of Phase 2 is a tool for default detection. This tool could easily be used to identify violations of already known defaults.

## 5.3 Chapter Summary

This chapter summarizes and discusses the results from both Phase 1 and Phase 2. The potential for future work in Ontological Semantic defaults is also discussed.

# REFERENCES

# REFERENCES

Appelt, D. E., & Kronfeld, A. (1987). A computational model of referring. IJCAI (Vol. 87, pp. 640-647).

Atifi, Hassan, Mandelcwajg, Sacha, & Marcoccia, Michel. (2011). The co-operative principle and computer-mediated communication: The maxim of quantity in newsgroup discussions. Language Sciences, *33*(2), 330-340.

Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, *53*(4), 742-753.

Baltes, P. (1995). Discourse reduction and ellipsis: A semantic theory of interpretation and recovery.

Beun, R., & Cremers, A. (1998). Object Reference in a Shared Domain of Conversation. Pragmatics & Cognition 6(1-2), 121-152 (1998).

Bishop, M., Engle, S., Peisert, S., Whalen, S., Gates, C., Probst, C., & Somayaji, A. (2008). We have met the enemy and he is us. *New Security Paradigms Proceedings of the 2008 Workshop*, 1-12.

Bradley, G., Sparks, B., & Nesdale, D. (2001). Doctor Communication Style and Patient Outcomes: Gender and Age as Moderators. *Journal of Applied Social Psychology*, *31*(8), 1749-1773.

Clark, H. H., & Marshall, C. R. (2002). Definite reference and mutual knowledge. *Psycholinguistics: Critical Concepts in Psychology*, *414*.

Cramton, C. D. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science*, *12*(3), 346-371.

Dale, R. (1992). Generating referring expressions: Building descriptions in a domain of objects and processes.

Dale, R., & Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expression.

De Marneffe, M. C., & Manning, C. D. (2008). The Stanford typed dependencies representation. *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation* (pp. 1-8). Association for Computational Linguistics.

Fan, X., Chen, X., & Hu, H. (2010, December). Utilizing background corpus and dictionary to calculate similarity between unknown words. *Information Science and Engineering (ICISE), 2010 2nd International Conference on* (pp. 1669-1672). IEEE.

Feldman, R., & Hirsh, H. (1997). Exploiting background information in knowledge discovery from text. *Journal of Intelligent Information Systems*, *9*(1), 83-97.

Francis, W. N., & Kucera, H. (1979). Brown corpus manual. Brown University.

Gibbs Jr., R. W. (1987). Mutual knowledge and the psychology of conversational inference. *Journal of Pragmatics*, 11(5), 561-588

Goodman, B. (1986). Reference identification and reference identification failure. *Computational Linguistics*, 12, 273-305.

Grice, H. P. (2013). logic and conversation. *The Semantics-Pragmatics Boundary in Philosophy*, 47.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the Web as Corpus. *Computational Linguistics, 29*(3), 333-347.

Krauss, R. M., & Fussell, S. R. (1990). Mutual knowledge and communicative effectiveness. *Intellectual teamwork: Social and technological foundations of cooperative work*, 111-146.

Krovetz, R. (2000). Viewing morphology as an inference process. *Artificial Intelligence*, *118*(1), 277-294.

Kulle, D, Kahana, E, & Kahana, B. (2012). Doctor-Patient communication: Gender makes a difference. *Psycho - Oncology, 21*, 86.

Magklaras, G. B., Furnell, S. M., & Brooke, P. J. (2006). Towards an insider threat prediction specification language. *Information Management & Computer Security*, *14*(4), 361-381.

Maree, M., & Belkhatir, M. (2013). Coupling semantic and statistical techniques for dynamically enrichin web ontologies. *Journal of Intelligent Information Systems*, *40*(3), 455–478. doi:10.1007/s10844-012-0233-4.

McShane, M., & Ebrary, Inc. (2005). A theory of ellipsis. Oxford ; New York: Oxford University Press.

Nirenburg, S., & Raskin, V. (2004). Ontological semantics (Language, speech, and communication). Cambridge, Mass.: MIT Press.

Novak, B. (2004). A survey of focused web crawling algorithms.

Paternotte, E., van Dulmen, S., van der Lee, N., Scherpbier, A. J., & Scheele, F. (2014). Factors influencing intercultural doctor–patient communication: A realist review. *Patient Education and Counseling*.

Porter, M. F. (1980). An algorithm for suffix stripping. Program, *14*(3), 130-137.

Prince, E. (1981). Toward a taxonomy of given-new information.

Reiter, E. (1991). A new model of lexical choice for nouns1. *Computational Intelligence, 7*(4), 240-251.

Raskin, V., Taylor, J. M., & Hempelmann, C. F. (2010, September). Ontological semantic technology for detecting insider threat and social engineering. *Proceedings of the 2010 workshop on New security paradigms* (pp. 115-128). ACM.

Ringenberg, T.R., Taylor, J.M., Springer, J.A., & Raskin, V. (2015). Towards identifying ontological semantic defaults with big data: Preliminary results. *Creative Content Technologies, 2015 International Conference.*

Ringenberg, T.R., Stuart, L.M., Taylor, J.M., & Raskin, V. (2015). Towards computer understanding of direct object defaults. Submitted for publication.

Schank, R., & Abelson, R. (1977). Scripts, plans, goals, and understanding : An inquiry into human knowledge structures (Artificial intelligence series (Hillsdale, N.J.)). Hillsdale, N.J. : New York: L. Erlbaum Associates ; distributed by the Halsted Press Division of John Wiley and Sons.

Soe Lai Phyue. (2013). Unknown Word Detection via Syntax Analyzer. *IAES International Journal of Artificial Intelligence* (IJ-AI), 2013, *2*(3).

Sperber, D., & Wilson, D. (1994). Outline of relevance theory. *Links and Letters*, 85-106.

Taylor, J., & Raskin, V. (2012). Understanding and structuring NL descriptions: The case of 101 animals. *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, 2638-2643.

Taylor, J.M., Raskin, V., & Hempelmann, C. (2011). From disambiguation failures to common-sense knowledge acquisition: A day in the life of an ontological semantic system. *Proceedings of the 2011IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 01, 186-190.

Taylor, J. M., Raskin, V., Hempelmann, C., & Attardo, S. (2010, October). An unintentional inference and ontological property defaults. *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on* (pp. 3333-3339). IEEE.

Taylor, J. M., Hempelmann, C., & Raskin, V. (2010). On an automatic acquisition toolbox for ontologies and lexicons in ontological semantics. *IC-AI* (pp. 863-869).

Teal, Cayla R., & Street, Richard L. (2008). Critical elements of culturally competent communication in the medical encounter: A review and model. *Social Science & Medicine*, *68*(3), 533-543.

Uson, R. M., & Perinan-Pascual, C. (2009). The anatomy of the lexicon within the framework of an NLP knowledge base. *Revista Española de Lingüística Aplicada*, 22.

Van der Sluis, I., Luz, S., Breitfuß, W., Ishizuka, M., & Prendinger, H. (2012). Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human Computer Studies*, *70*(9),611-629.

Viethen, J., & Dale, R. (2006). Algorithms for generating referring expressions: do they do what people do?. *Proceedings of the Fourth International Natural Language Generation Conference* (pp. 63-70). Association for Computational Linguistics.

Wikipedia for Schools. (n.d.). Retrieved 2015, from http://schools-wikipedia.org/

Wilson, D., & Sperber, D. (2002). Truthfulness and relevance. *Mind*, *111*(443), 583-632.

Zhang, J., Zhai, F., & Zong, C. (2013). A substitution-translation-restoration framework for handling unknown words in statistical machine translation. *Journal of Computer Science and Technology*, *28*(5), 907-918.

# APPENDICES

# Appendix A

## Sample Data from Phase 1

Table A. 1 *Wikipedia With Adjective Only Sample*

| Verb | Noun | Adjective |
|---|---|---|
| accept | ceasefire | de |
| accept | certainty | alternative |
| accept | fact | unpleasant |
| accept | institutions | existing |
| accept | Laos | neutral |
| accept | man | outstanding |
| accept | miracle | greater |
| accept | object | entering |
| accept | order | misshapen |
| accept | results | chief |
| accept | standard | double |
| accept | timetable | new |
| accepted | bowl | silver |
| accepted | planning | longrange |
| accepted | proposals | such |
| accepted | style | Geometric |
| accepting | faith | Christian |
| achieve | ambitions | legitimate |
| achieve | cooperation | perceptive |
| achieve | objectives | longrange |
| achieve | recovery | economic |
| achieve | result | just |
| achieve | salvation | personal |
| achieve | stature | Christian |
| achieved | record | brilliant |
| achieved | state | high |
| achievements | physics | atomic |
| achieves | government | democratic |
| achieving | objectives | limited |
| act | issues | sore |
| act | way | same |
| add | book | more |
| add | book | new |
| add | color | much |

# Appendix B

## Sample of Wikipedia Data from Phase 2

Table B. 1 *Wikipedia Data using Adjective-Only Method*

| Verb | Verb | | |
|---|---|---|---|
| Infinitive | Form | Noun | Adjective |
| mark | mark | intimacy | mutual |
| mark | mark | kind | same |
| mark | marks | latitude | northernmost |
| mark | marks | latitude | southerly |
| mark | marks | latitude | southernmost |
| mark | mark | latitudes | northernmost |
| mark | marking | legislation | federal |
| mark | marking | legislation | first |
| mark | marked | lento | Andante |
| mark | marked | letter | last |
| mark | marking | life | contemplative |
| mark | marked | life | everyday |
| mark | marked | life | later |
| mark | marked | limes | northern |
| mark | marked | limes | northern |
| mark | marks | limit | eastern |
| mark | marked | limit | navigable |
| mark | marks | limit | northernmost |
| mark | marks | limit | northernmost |
| mark | marks | limit | southern |
| mark | marked | limit | southern |
| mark | marks | limit | upstream |
| mark | marked | limit | western |
| mark | mark | limit | western |
| mark | marked | limits | distinct |
| mark | marking | line | central |
| mark | mark | line | current |
| mark | mark | line | dividing |
| mark | mark | line | east-west |
| mark | marking | link-up | first |
| mark | marking | link-up | such |
| mark | marks | location | _ |
| mark | marks | location | approximate |
| mark | mark | location | approximate |

Table B. 2 *Wikipedia Data using All-Modifiers Method*

| Verb<br>Infinitive | Verb Form | Noun | Adjective | Other Modifier |
|---|---|---|---|---|
| form | form | stretch | deadliest | |
| form | form | stretch | longest | |
| form | form | stretch | single | |
| form | form | strictly | | defined |
| form | form | strictly | single | |
| form | forming | string | closed | |
| form | form | string | complete | |
| form | form | strip | | entitled |
| form | form | strip | new | |
| form | forming | Strip | | Cotai |
| form | forming | stroke | short | |
| form | forms | structure | | generates |
| form | forming | structure | | market |
| form | forming | structure | | known |
| form | forming | structure | | known |
| form | forms | structure | | known |
| form | form | structure | | called |
| form | form | structure | | Management |
| form | form | structure | | support |
| form | forming | structure | | requires |
| form | form | structure | | rope |
| form | form | structure | | planar |
| form | form | structure | | ring |
| form | form | structure | | is |
| form | form | structure | | crystal |
| form | forming | structure | | known |
| form | form | structure | | differentiates |
| form | form | structure | | jaw |
| form | form | structure | | stressed |
| form | forms | structure | | known |
| form | form | structure | | enable |
| form | form | structure | | trade |
| form | form | structure | | union |
| form | form | structure | | crystal |
| form | form | structure | | have |
| form | forms | structure | | capital |
| form | forms | structure | | company |
| form | form | structure | | wall |
| form | form | structure | | crystal |

# Appendix C

## Sample of Brown Corpus Data from Phase 2

Table C. 1 *Brown Corpus Data with Adjective-Only Method*

| Verb Infinitive | Verb Form | Noun | Adjective |
|---|---|---|---|
| accept | accepted | bowl | silver |
| accept | accept | cease-fire | de |
| accept | accept | certainty | alternative |
| accept | accepting | faith | Christian |
| accept | accept | Laos | neutral |
| accept | accept | miracle | greater |
| accept | accept | object | entering |
| accept | accept | order | misshapen |
| accept | accept | timetable | new |
| achieve | achieve | cooperation | perceptive |
| achieve | achieves | government | democratic |
| achieve | achieving | objectives | limited |
| achieve | achieve | stature | Christian |
| act | act | issues | sore |
| add | add | book | bad |
| add | add | book | more |
| add | add | conception | second |
| add | add | interest | geological |
| add | add | members | new |
| add | add | mustard | prepared |
| add | add | note | colorful |
| add | add | note | decorative |
| add | add | note | do-it-yourself |
| add | add | note | human |
| add | added | pars | straight |
| add | adding | pieces | small |
| add | add | reform | more |
| add | add | reform | practical |
| add | add | tablespoons | several |
| add | add | touch | exciting |
| allow | allow | autonomy | greater |
| allow | allow | collection | further |
| allow | allow | contests | endurance |
| allow | allow | contests | underwater |

Table C. 2 *Brown Corpus Data using All-Modifiers Method*

| Verb Form | Noun | Adjective | Other Modifier |
|---|---|---|---|
| accepted | bowl | | his |
| accepted | bowl | silver | |
| accept | cease-fire | de | |
| accept | certainty | alternative | |
| accept | concessions | | seniority |
| accepting | faith | Christian | |
| accept | findings | | Freud |
| accept | Laos | neutral | |
| accept | miracle | greater | |
| accept | object | entering | |
| accept | order | | challenges |
| accept | order | misshapen | |
| accept | sacrifice | | his |
| accept | timetable | new | |
| achieve | cooperation | perceptive | |
| achieved | following | | such |
| achieves | government | democratic | |
| achieving | objectives | limited | |
| achieve | stature | Christian | |
| act | issues | | plague |
| act | issues | sore | |
| adding | bit | | his |
| add | book | | one |
| add | book | bad | |
| add | book | more | |
| add | conception | | his |
| add | conception | second | |
| add | contribution | | only |
| add | contribution | | one |
| add | cup | | half |
| add | inhibitor | | rust |
| add | interest | geological | |
| add | members | | three |
| add | members | new | |
| add | mustard | prepared | |
| add | note | | interest |
| add | note | colorful | |
| add | note | decorative | |