Building Representations from Natural Language

by

Mark J. Seifter

S.B. Bachelor of Science in Computer Science and Engineering
Massachusetts Institute of Technology, 2007

Submitted to the Department of Electrical Engineering and Computer Science

In Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

At the Massachusetts Institute of Technology

August 2007

Author: ____
Department of Electrical Engineering and Computer Science
August 31, 2007

Certified by: ____
Patrick H. Winston, Ford Professor of Artificial Intelligence and Computer Science
Thesis Supervisor

Accepted by:____
Arthur C. Smith
Professor of Electrical Engineering
Chairman, Department Committee for Graduate Theses

Building Representations from Natural Language
by
Mark J. Seifter

Submitted to the
Department of Electrical Engineering and Computer Science

August 31, 2007

In Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

In this thesis, I describe a system I built that produces instantiated representations from descriptions embedded in natural language. For example, in the sentence 'The girl walked to the table', my system produces a description of movement along a path (the girl moves on a path to the table), instantiating a general purpose trajectory representation that models movement along a path.

I demonstrate that descriptions found by my system enable the imagining of an entire inner world, transforming sentences into three-dimensional graphical descriptions of action. By building action descriptions from ordinary language, I illustrate the gains we can make by exploiting the connection between language and thought.

I assert that a small set of simple representations should be able to provide powerful coverage of human expression through natural language.

In particular, I examine the sorts of representations that are common in the Wall Street Journal from the Penn Treebank, providing a counterpoint for the many other sorts of analyses of the Penn Treebank in other work. Then, I turn to recognized experts in provoking our imaginations with words, using my system to examine the work of four great authors to uncover commonalities and differences in their styles from the perspective of the way they make representational choices in their work.

Thesis Supervisor: Patrick Henry Winston
Title: Ford Professor of Artificial Intelligence and Computer Science

Acknowledgements

I would like to thank fellow members of the Neo-Bridge project Harold Cooper and Diana Moore and the many other members of the Genesis group for the part they played in my research, whether it be designing components to interface with Span, looking over my thesis, or just letting me come to them to bounce ideas.

I would like to especially thank Patrick Winston and Mark Finlayson for letting me explore the big ideas of Artificial Intelligence and pursue my own projects. And, of course, I thank them for their tireless efforts at revising this thesis into a readable, cohesive whole.

Finally, I would like to thank my parents for their tireless focus on my education. I wouldn't be here without your support.

Contents

Chapter 1

Introduction

1.1 The Vision

I believe that in order to create a functioning artificial intelligence, we must first understand human intelligence and what makes our intelligence unique. Forays into the understanding of intelligence are doomed to failure until we take into account the way tightly coupled loops between senses and language shape our thought and form our responses. Therefore, if we are to discover the nature of intelligence, we need to think about what we can learn about intelligence and thought from language. We need to uncover the hidden clues in our speech that underlie representations of the world in our head. In this thesis, I explore the process of understanding what we think from what we say.

1.2 Motivating Example

To motivate my vision, imagine some time in the near future when a human is conversing with a computer system designed to build cognitive representations and think about any natural language inputs it receives. We might see a simple interaction like this one:

Human:“Did you see that? The man just flew to the tree?”
System: “That is unusual. A man usually cannot fly. How did the man fly to the tree?”
Human: “Oh, the man flew because he was wearing a jetpack.”
System: “I don't know what a jetpack is.”
Human: “It is a type of equipment.”
System: “Okay, I do know what equipment is, but usually wearing equipment does not let a man fly. I assume a jetpack is different from normal equipment because wearing a jetpack allows a man to fly?”
Human: “That is correct.”
System: “So noted. Wasn't there a wall between the man and the tree? I assume the man flew over the wall?”
Human: “Yes he did.”

In this interaction, the computer has built an image of the world for itself based on the conversation with the human, and it asks questions to fill in the inconsistencies it finds with its own knowledge of the way things interact. In this example, the computer talking to the human is much like a child playing a game of make-believe or reading a story, as it uses language to imagine an entire world that it cannot experience directly, and it reasons and learns about that world based on the representations and imaginings it creates. A system like this has taken an important step towards the sonnet-writing machine that debates the line 'Shall I compare thee to a summer's day' in Alan Turing's famous human/computer dialogue (Turing 1963).

But how can we create such a system? In the past, the closest we have is a modified theorem prover that can take natural language inputs that correspond directly to logical statements and tell you if they contradict. For instance, such a system might scan “Peter loves every girl.” “Abby is a girl.” “Peter does not love Abby” and state a contradiction. But even though the system has come to the correct conclusion, This system has not really learned anything about the world—it simply converted words into logic and deduced the obvious contradiction. In order to discover what we think from what we say, we need to avoid a direct conversion into logic. Instead, I will focus on building cognitive representations from natural language and understanding the way we imagine and hallucinate.

1.3 Steps towards the Vision

In this thesis, I describe a system I built that produces instantiated representations from descriptions embedded in natural language. For example, in the sentence ‘The girl walked to the table’, my system produces a description of movement along a path (the girl moves on a path to the table), instantiating a general purpose trajectory representation that models movement along a path.

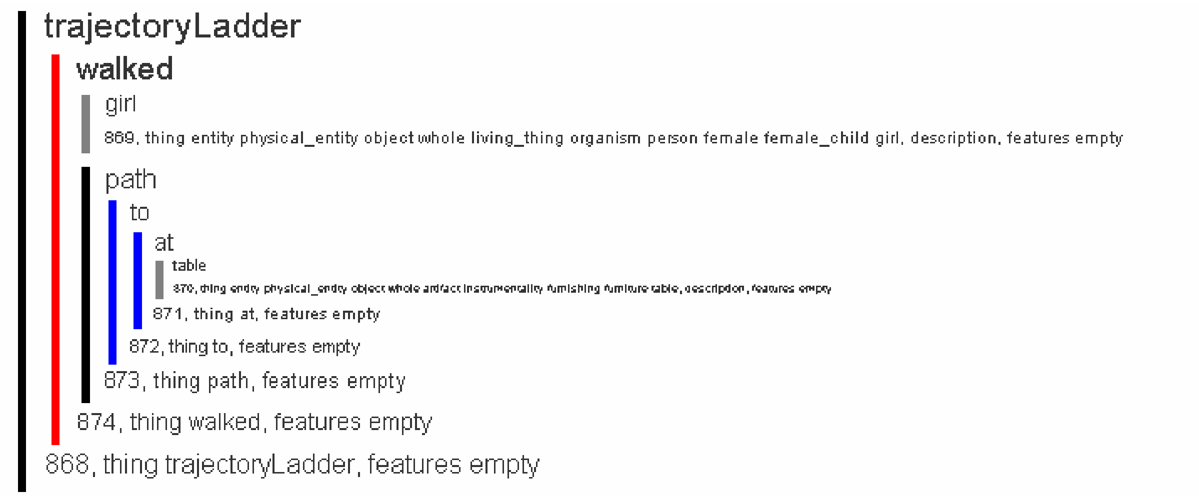

Fig 3.1: A sample Trajectory for the sentence “The girl walked to the table”.

I demonstrate that descriptions found by my system enable the imagining of an entire inner world, transforming sentences into three-dimensional graphical descriptions of action. By building action descriptions from ordinary language, I illustrate the gains we can make by exploiting the connection between language and thought.

I assert that a small set of simple representations should be able to provide powerful coverage of human expression through natural language.

In particular, I examine the sorts of representations that are common in the Wall Street Journal from the Penn Treebank, providing a counterpoint for the many other sorts of analyses of the Penn Treebank in other work. Then, I turn to recognized experts in provoking our imaginations with words, using my system to examine the work of four great authors to uncover commonalities and differences in their styles from the perspective of the way they make representational choices in their work.

Through my analysis, I discover some interesting nuances in the authors’ writing styles. Dickens changes the way he describes his world substantially between his writings in *David Copperfield* and *Oliver Twist*. Jane Austen’s work in *Pride and Prejudice* and *Sense and Sensibility* is nearly identical from a representational point of view and has several distinctive aspects, including frequent use of descriptions that ascribe an attribute. For example, the sentence “Mrs Dashwood was surprised only for a moment at seeing him” has an embedded description—Mrs. Dashwood was surprised.

Chapter 2

Vignettes in Language and Thought

For over half a century, research in the field of Artificial Intelligence has sought to produce a thinking machine, the Holy Grail of AI predicted by Alan Turing in 1936. In so doing, unfortunately, much of the field has lost sight of the fact that there are other fields that have been trying to explore human intelligence for millennia. In abstracting everything away to a world of logic and hard-coded rules as in a symbolic cognitive architecture or rejecting all representations in favour of pure task-based stimulus response as in the subsumption architecture, many in the field of Artificial Intelligence have overlooked the important aspect of intelligence inherent in the integration of senses and reasoning. We think with our hands, we think with our eyes, we think with our mouths.

The fact that language and thought are inextricably intertwined does not come at all as a surprise to linguists, cognitive scientists, and others who have studied intelligence in humans. Even as far back as the sophists of Ancient Greek, humans have puzzled over how language influences thought. They catalogued the ways in which language could be used to influence the mind, developing the school of rhetoric.

The Sapir-Whorf Hypothesis in linguistics (Sapir 1929, Whorf 1956) formalised the concept that language and thought were inextricably intertwined. In literature, George Orwell's *1984* discussed the creation of a new language, Newspeak, that would make disloyal thoughts impossible “It was intended that when Newspeak had been adopted once and for all and Oldspeak forgotten, a heretical thought--that is, a thought diverging from the principles of Ingsoc--should be literally unthinkable, at least as far as thought is dependent on words.” (Orwell 1948) Modern philosopher Ludwig Wittgenstein wrote in his *Tractatus Logico Philosophicus* “The limits of my language indicate the limits of my world” (Wittgenstein 1966).

Many enterprising scientists have performed experiments in an attempt to understand the role of language in our thought. Behavioural economics has studied the effects of framing and how the language of the choices can cause drastic differences in the selection of equivalent options (Tversky 1981). For instance, in the 'Asian disease' scenario where an untreated disease will kill 600 people, subjects were 72% likely to prefer “200 people will be saved” to “there is a one-third probability that 600 people will be saved, and a two-thirds probability that no people will be saved ”, but they were 78% likely to prefer “there is a one-third probability that nobody will die, and a two-third probability that 600 people will die” to “400 people will die”.

In an attempt to study the way language influences perception, Borodisky performed an anthropological study of an island culture with no word for relative directions. Instead, the culture used only 'North', 'East', 'South', and 'West' for all directional descriptions. In addition to having an inherent sense of north, the people of this culture displayed an intriguing tendency towards their view of an imagined pictorial representation of the passage of time. If you asked an American child to order five pictures that make up a story from first to last, the child would invariably place the pictures in order from left to right. The children of this culture, however, always placed the pictures from east to west, like the rising and setting sun. It didn't matter which direction the experimenter initially placed the child. The child would arrange them vertically or horizontally, left-to-right, right-to-left, up-to-down, or down-to-up, whichever was required to arrange them east to west. This experiment suggests that the way these people imagine and represent a series of events and the passage of time is intrinsically linked to the way they express those concepts in their language. If the islanders possessed a linguistic concept of relative direction, they might have arranged the story in a standardized relative direction, for instance left to right, much like people in other cultures.

Elizabeth Spelke's dual task studies have also strongly indicated that the use of language is what allows humans to combine modalities and outperform other animals at cognitive tasks (Spelke 1999). In Spelke's experiments, she placed an object in one corner of a rectangular room while a subject observed. She then disoriented the subject and tasked the subject to locate the object. Because of the geometry of the room, there were two pairs of geometrically equivalent corners (long side left or short side left). Additionally, one wall of the room was painted blue. Rats and young children searched the two geometrically equivalent corners with equal probability, unable to combine the colour and the geometry to find the only solution, even though in another experiment, both rats and infants could use the colour cue alone. Adults and children, starting at the age when the children learned how to use locational words like 'left', 'up', and 'next to' generatively in their own speech, were able to locate the correct corner by combining the geometric and colour cues. Interestingly, when they were engaged in a verbal shadowing task to distract their linguistic processor, adult humans performed just as poorly as the rats.

Chapter 3

A System of Representation

3.1 The Span System Basics

I created the Span (also known as Neo-Bridgespeak) system in order to perform experiments on language and thought using any sort of unformatted natural language as an input. The idea to create a bridge between language and representations is not a new one—in fact, Patrick Winston and the Genesis Group's original Bridge system was able to do so for an extremely restricted subset of words that the system knew beforehand (Bender, 2001; Bonawitz, 2003; Larson, 2003; Molnar, 2001; Shadadi, 2003).

What makes Span unique is that it combines insight into cognitive representations of the world with the language capabilities of state-of-the-art statistical parsers, a blend that allows robust performance on a wide variety of expressions. Whereas Bridge could only handle words that it knew in advance that were arranged together in very specific ways, Span is able to detect the use of any of the representations it knows, even within highly complex sentence structures.

The only limit to Span is what can be parsed by the statistical parser component, and it is built in a modular fashion such that the parser can be switched out with very little effort. By default, the Span system uses the Stanford parser, available free online at http://www-nlp.stanford.edu/downloads/lex-parser.shtml, but it works perfectly well with any parser that produces parse trees with the standard part-of-speech tags used in the Penn Treebank.

3.2 How It Works

Span works by taking an input in natural language and feeding it to the parser component in order to retrieve a parse tree. Once it has obtained the parse tree, the system searches the tree for a substructure that might indicate the presence of one of the representations it knows, using a regular expression search for trees (or 'tregex'). Teaching the system a new representation is as simple as adding the representation to the list that the system checks when it sees a new tree. Sometimes, a complex or compound sentence may use several different representations, and Span is able to find as many as it can uncover in the tree structure from the parser.

In addition to the basic pattern matching, Span performs further checking and reasoning specific to each rule encoded in Span for a representation. For instance, “NounX Verbed NounY” will very clearly contain an important description, but based on the structure alone, there is not enough information to disambiguate between several possibilities. By reasoning based on word knowledge, the system is able to determine which representation actually applies in any given substructure of the parse tree.

3.3 Building a Tregex Match

In order to create the tregex pattern for a given representation, I tested many sentences containing the given representation on the Stanford parser and looked carefully at the output parse trees.

Although the basic patterns were usually simple, creating the specifics was often nuanced by exceptions in parse tree structure (for instance, it was not always sufficient to expect a noun phrase to directly lead into the object of a preposition or a verb phrase—sometimes the parser would insert additional layers of noun phrases between the top of the expression and the actual representation. Thus, rather than detail a highly specific and regimented pattern that had poor coverage of expressions invoking the representation, I used tregex expressions to tell the system that while there may not be a noun phrase leading to what it wants to find right now, it needs to check to see if that appears somewhere down the line, possibly below multiple levels of redundant noun phrases. This relation is called 'dominates' in tregex.

The good thing about English as a language is that the key word in a given phrase is always on the left (excepting perhaps noun phrases, but some linguists quibble that the article is the key word in a noun phrase anyway), so it is easy enough to search for a match of key words by using tregex's 'first child' operation, and then the 'sibling of' operation to find the other parts of the pattern in that same phrase.

Tregex also allowed me to use an assignment operation in order to assign a name to certain portions of the pattern, thus making retrieval of important parts of the representation extremely simple.

So here's an example: A Relation will start out at the sentence level, and it will dominate, eventually, a noun phrase (labeled as the Subject Noun Phrase) which will be a sibling of a verb phrase. The verb phrase will dominate a verb (labeled as the Verb) and then another noun phrase whose last constituent will be a noun (labeled as the Object).

Of course, this pattern is insufficient to determine which representation I have found—the same pattern also matches the Implied Transitive Trajectory. Thus, after searching the pattern, the Representation matcher performs several checks on the resulting subtree to determine the correct representation. Specifically, the Implied Transitive Trajectory pattern checks to see if it knows any implied trajectories for the given verb. If so, it creates the implied trajectory, and if not, it returns a failure and allows the Relation pattern to try instead.

3.4 Span's Representations

The basic representations that Span understands are Is-A, Is-JJ, Is-Superlative, Is-Possessive, Trajectory, Relation, and Transition. Even though these representations are simple, they can be used to reason about the world with a wide degree of coverage.

The Is-A representation expresses the fact that one entity, the subject, is a member of another class of entities, and it can be used to connect knowledge from the second class of entities to the subject. For instance, when the human in my motivating example tells the computer that a jetpack is a type of equipment, the computer can now use its knowledge of equipment to think and reason about the jetpack. Is-A may be simple, but it is crucial for building equivalence classes. Sometimes the best way to quickly bring someone else to understand something is by equating it to something else (e.g. "Professor Smith is a cross between Severus Snape and Dr. Frankenstein"). Additionally, Is-A representations can be used to express metaphor (e.g. "Juliet is the sun" rather than "The cat is an animal"). When Is-A is in its simplest form, however, it expresses a hyponym/hypernym relationship between the subject of the sentence and the predicative nominative. In fact, Marti Hearst (Hearst 1992) used such contextual clues to automatically build a large-scale dictionary of hyponym/hypernym relationships from simple natural language text. A more finely-tuned future study might consider the distance between two nouns in the existing hypernym structure in an attempt to predict when metaphor is in use (for instance, "Juliet" and "sun" would be much further away in the hypernym structure than "cat" and "animal").

The Is-JJ representation handles the case when an entity, the subject, is stated to possess a certain property. The Span system uses Thread Memory (Greenblatt 1979) to store properties—each object possesses a 'Properties' thread, and an Is-JJ representation can be used to add additional properties, though they can also be added by placing the adjectives in front of the noun as in “The big red dog”. In the motivating example, the computer has taken this idea one step further and added the 'wearer-can-fly' property to the object 'jetpack' in order to distinguish it from other types of equipment.

The Is-Superlative representation stores the knowledge that one entity, the subject, possesses some quality in a greater amount or capacity than another entity. Is-Superlative allows the system to develop a sense of how objects relate to each other in certain qualities. With a good number of Is-Superlative relations, a thinking machine could begin to build a partial order and ask the right questions about new data it receives at a later time. For instance, if a machine has received a large amount of Is-Superlative data about the heights of buildings, and a human tells it “Hey, wow! Tokyo Tower is really tall!” the machine might reply, “Interesting. But is it taller than the Sears Tower?”

The Is-Possessive representation covers the situation where one entity acts as another type of entity for a third possessor entity. For instance, “The penguin is the bear's food”. This situation is significantly more complex than the simpler use of the possessive in a sentence like “The chef's food is delicious”. In the simpler sentence, the possessive 'chef's' is just a property of food, and we can reason about this sentence perfectly well by just using the Is-JJ rule and adding in the property 'chef's'. But for “The penguin is the bear's food”, if we tried to just add 'bear's' as a property, we would end up with the statement that penguin is a type of food, which is not very helpful for reasoning. Instead, the Span system builds a 'food' relation between penguin and bear.

The Trajectory representation formalises movement along a path, either physical or metaphorical (so a trajectory covers both 'The man ran towards the woman' and 'Iraq moved towards Democracy'). Trajectories were studied carefully by Ray Jackendoff (Jackendoff 1983), and the Trajectory representation expresses useful information about the movement in the trajectory based on the use of prepositions. For instance, two major types of trajectory are 'open' and 'closed', where an open trajectory heads towards the destination but doesn't reach it, while a closed trajectory reaches its destination). By analysing prepositions from the trajectories, it is easy to determine whether the trajectory is open or closed (e.g. 'towards' would be open, and 'to' would be closed).

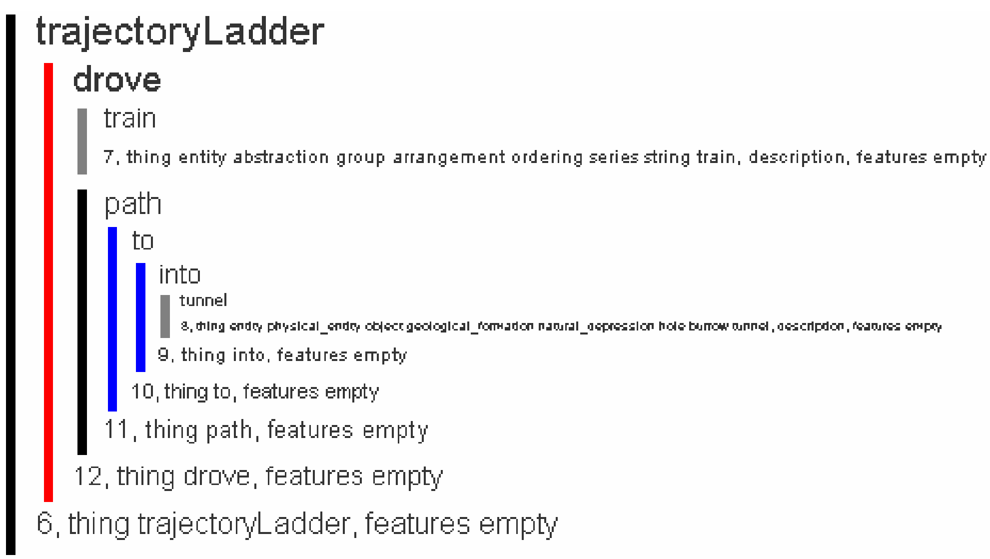

Fig 3.1: A sample Trajectory for the sentence “The train drove into the tunnel”.

Additionally, certain verbs have implied trajectories hidden within them. For instance, in the sentence “John ascended the cliff”, it is implied that John moved to the top of the cliff. In fact, the verb ascend contains the trajectory within it. Beth Levin categorised thousands of verbs in her

English Verb Classes and Alternations: A Preliminary Investigation, (Levin 1993), and the Span system searches all of these verbs for hidden trajectories when it encounters them.

Relation representations express a wide variety of active and passive relationships between two entities. A Relation covers both “Macbeth killed Duncan” and “Romeo loves Juliet”, storing the knowledge in the familiar form Killed(Macbeth, Duncan) and Loves(Romeo,Juliet). By storing these relations when it finds them in a natural language text, a system built on top of Span could easily answer questions like “Who killed Duncan?” or “Who does Romeo love?” by accessing its relational knowledge. Such an approach has been applied successfully by Boris Katz (Katz 1997)

Transitions are a simple but powerful representation created by Gary Borchardt (Borchardt 1994). A Transition details a change, appearance, disappearance, increase, or decrease, or the lack of any of the above. Many complicated actions can be broken down into component transitions (indeed, the conversion between trajectories and transitions is highly useful in creating an Imaginer), and sometimes the entire thrust of a sentence is a single transition. This is often the case for intransitive verbs that do not have a hidden trajectory, so for instance “Thus, with a kiss, I die” contains 'I die', which is best modeled as the transition 'disappear'.

Chapter 4

Imagining a Scene

4.1 A Bridge to Vision

One immediate and provocative use of the Span system is to imagine and render three-dimensional graphical models based on the descriptions that the system builds from natural language. Thanks to graphics work done by Harold Cooper, we have expanded the Span system into an entire Neobridge system that descends from Winston's Bridge system.

The Neobridge system focuses mainly on the most physically demonstrative descriptions, which come from the Trajectory representation. Using the Span system as its core, it can receive any input of natural language and output a three dimensional scene of the trajectory encoded in the sentence. For instance, “The bird flew to the top of the tree.”

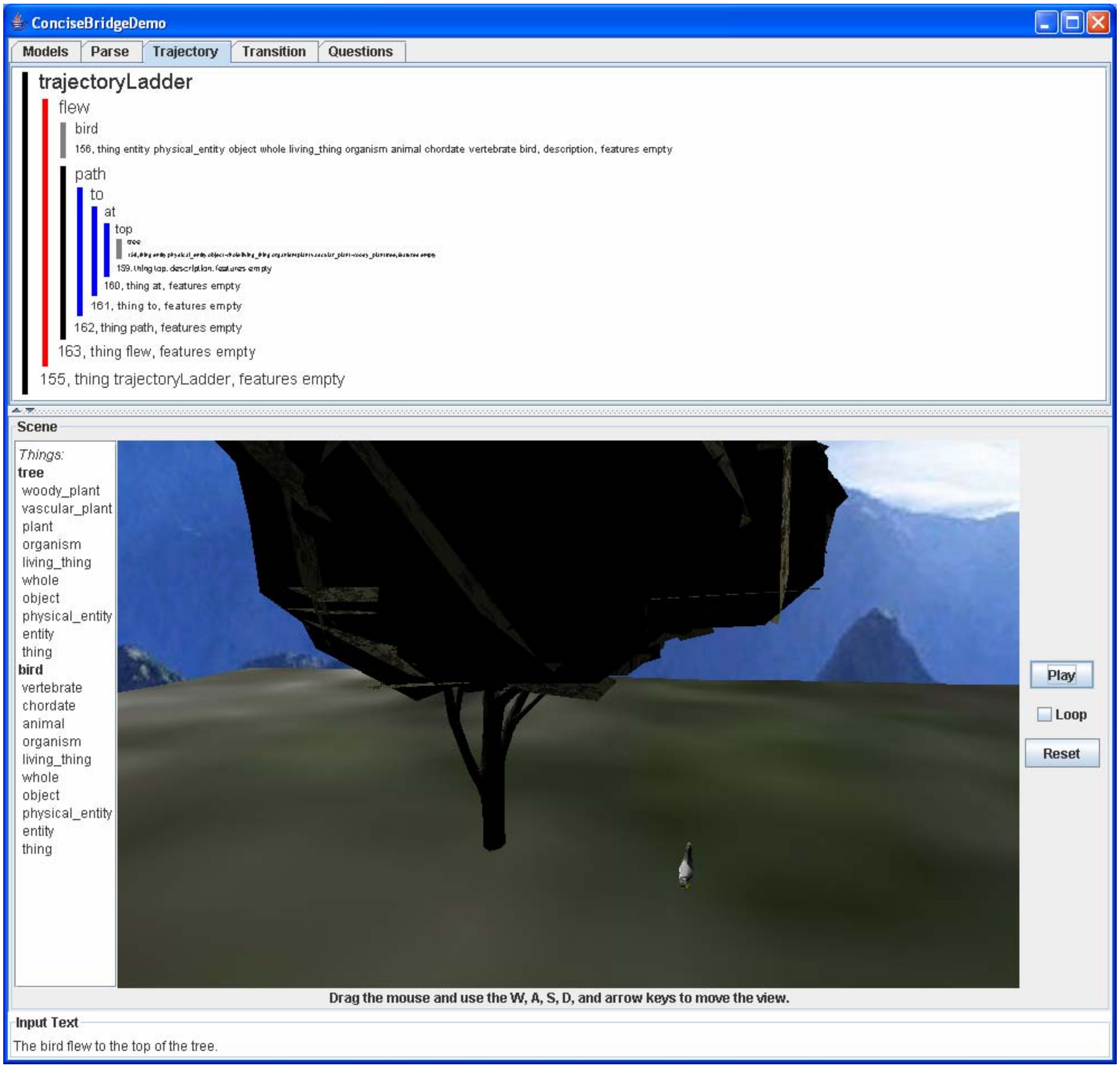

Fig 4.1: Three-dimensional model for the sentence "The bird flew to the top of the tree". The bird begins on the ground.

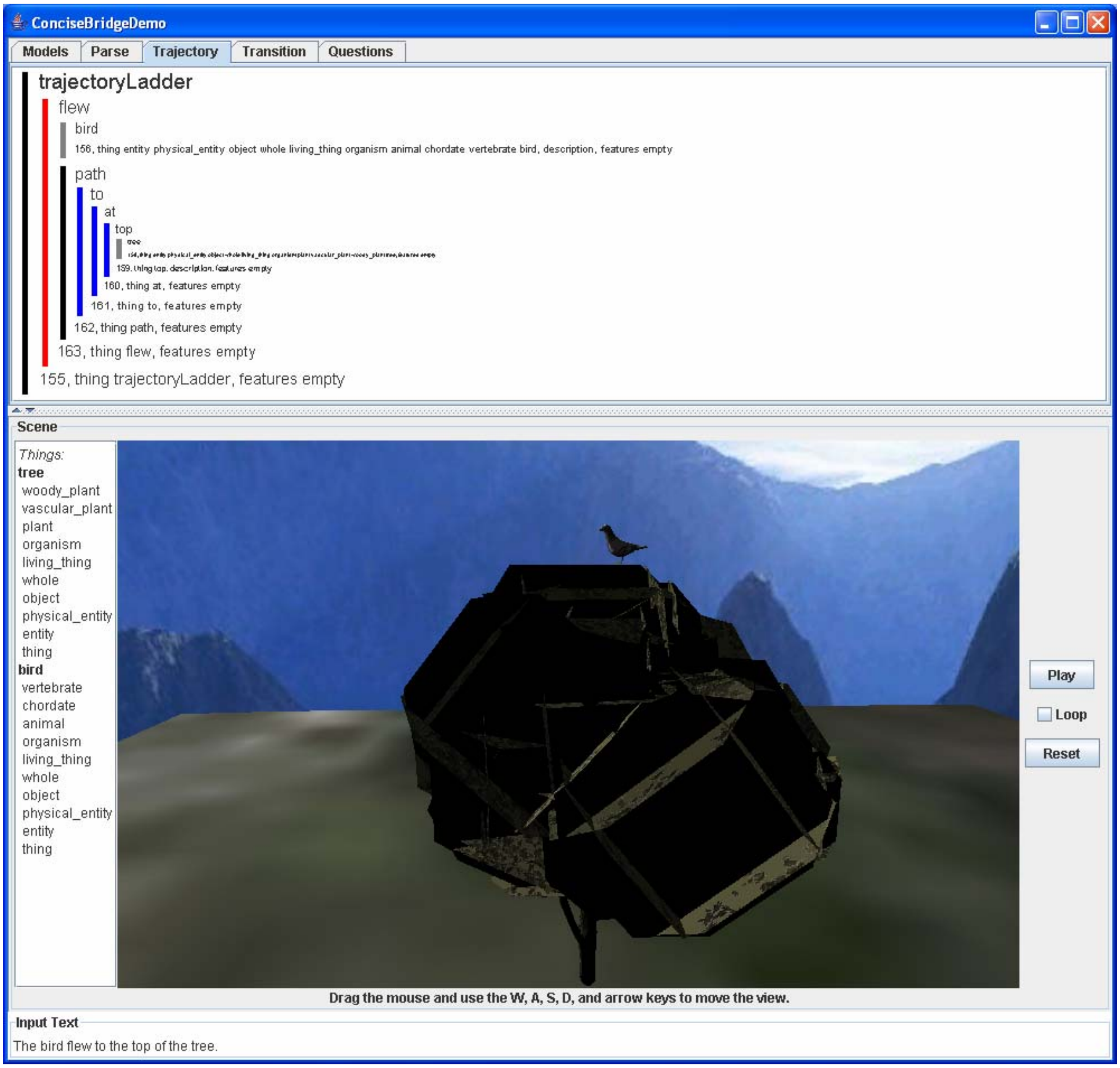

Fig 4.2: Three-dimensional model for the sentence "The bird flew to the top of the tree". The bird reaches the top of the tree.

The system can convert even complex trajectories like "The bird flew from the top of the table to the top of the tree into the trash can".

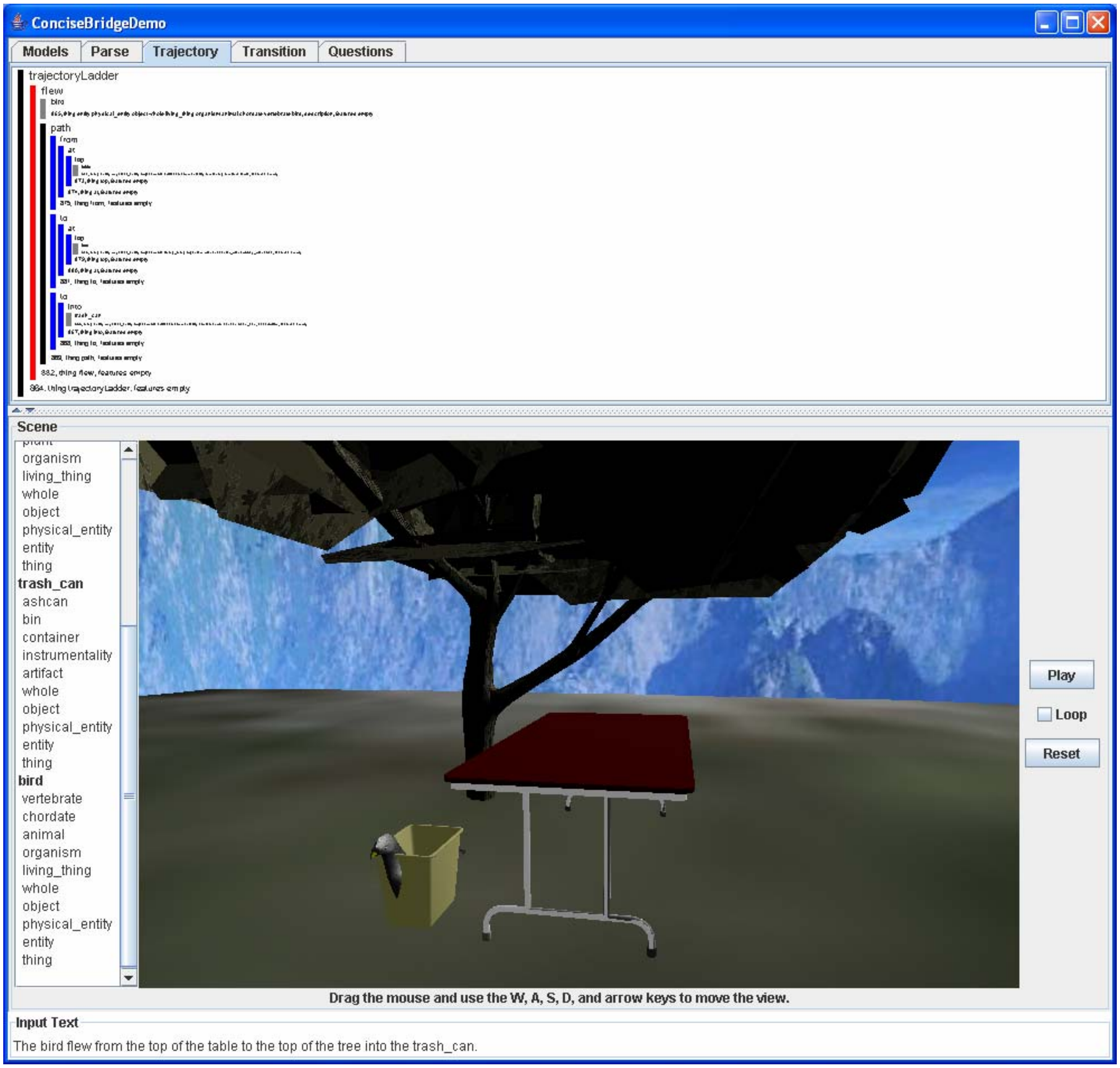

Fig 4.2: "The bird flew from the top of the table to the top of the tree into the trash can". The bird has reached its end goal and is inside the trash can.

The Neobridge system is able to call up the best images it knows to handle any given scene, even if it does not have images of the exact nouns that the user entered in the sentence. It does so by using a hypernym search through wordnet, beginning with the nouns in the input sentence and climbing up from hypernym to hypernym until it finds a word it knows. So for instance, while it may not know what an ibis is, it eventually finds the hypernym 'bird', and it does know what a bird is. If the system has no images for any hypernym, it will eventually reach the universal hypernym 'Thing' and just display a grey blob, as in "The country moves towards democracy".

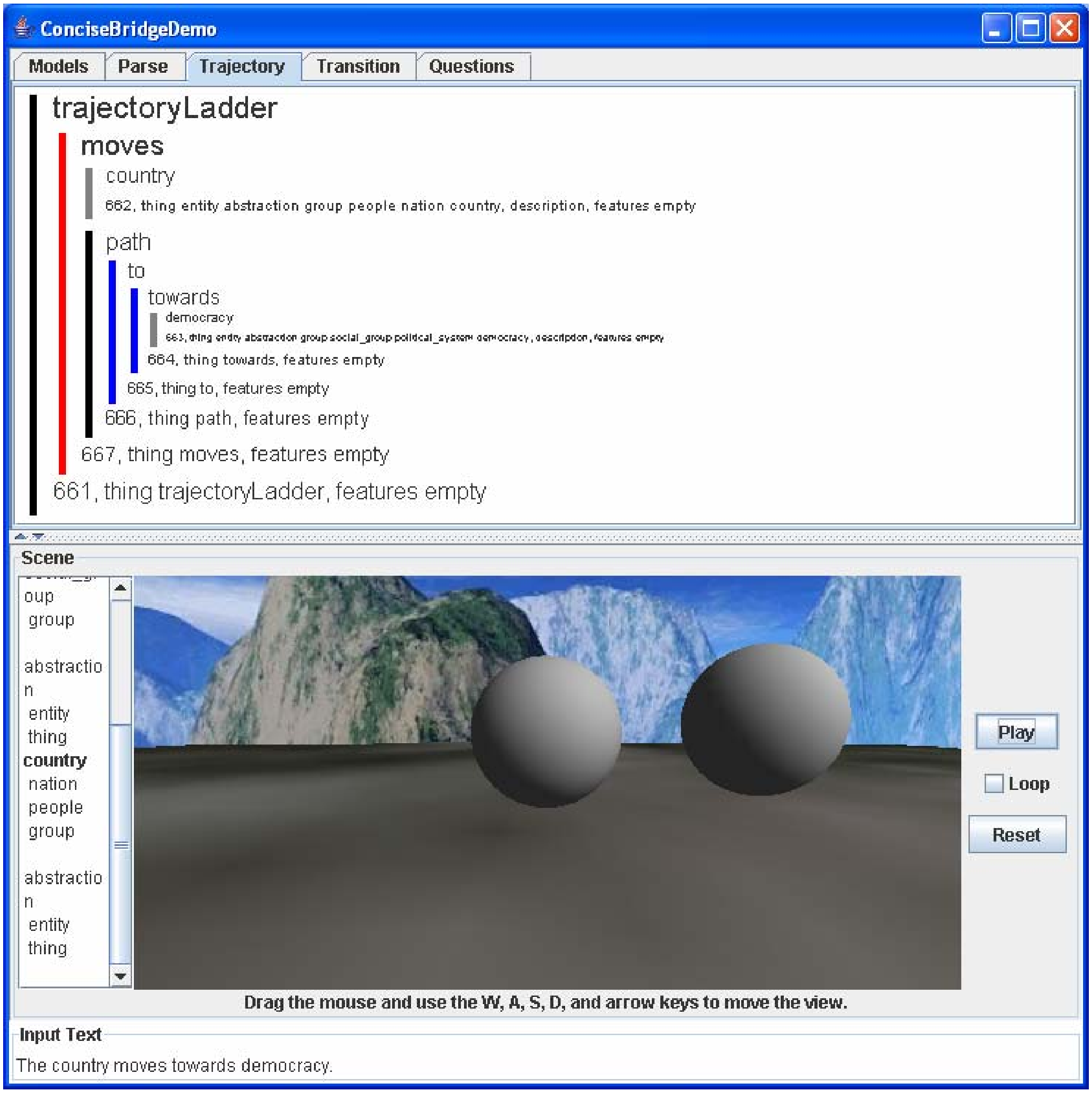

Fig 4.2: "The country moves towards democracy". The two entities with unknown pictures are represented as blobs.

4.2 Unimagining and Reasoning

Once the scene is created, the system then 'unimagines' the three-dimensional visual description into a series of Borchardt transitions, thus allowing Neobridge to use its spatial and visual components to discover new information that was not available directly from the language alone but was instead implied. Extracted information includes contact between objects and their speeds.

A human child is able to read the statement "John kissed Mary" and then correctly answer the question "Did John touch Mary?". Similarly, the system can use its visual and linguistic knowledge, stored in the unimagined Transition Space, to solve the same types of questions, questions that have been often used as examples of something that is easy for a human but devilishly difficult for a computer without a very specific self-defeating proclamation by the system's creator that kissing implies touching. The system also has a question-answering component, which allows comprehensive questions about the scene based on the transitions that take place.

Chapter 5

Analysing the Penn Treebank

Once I completed my system, I first decided to run it on the entire Penn Treebank, which consists of 50,000 sentences from the Wall Street Journal corpus. This would be a test of the representations my system could find in a large number of natural language inputs, though of course because the source was always from Wall Street Journal articles, I can make no claim that the results are representative of the average English sentence. Because the Penn Treebank is widely used in natural language studies, however, I felt that this would be an excellent common starting point for my research with my system.

5.1 Hypothesis

Before I started, I thought about what I expected from the Penn Treebank. Because it came from a newspaper, I expected a goodly number of declarative statements using Is-A or Is-JJ representations, with a large number of relations as the Wall Street Journal reports on the relationships between entities (things like “AT&T bought T-Mobile” or “Pierre Vinken led the board of directors”), with a few trajectories for the changing world of stocks (e.g. “The DOW increased today”) and a small number of superlatives (“The US Dollar is stronger than the Russian rubel”) and possessives (“Kenneth Lay is Enron's CEO”)

5.2 Results

Out of 50,000 sentences, 136 were unparsable by my system, so only about .27% of the sentences were thrown out. The raw numbers for each representation are as follows:

Total Sentences	49864
Is-A	2035
Is-JJ	2338
Superlative	48
Of Possessive	44
Apostrophe Possessive	21
Complex Trajectory	2636
Simple Trajectory	8207
Implied Transitive Trajectory	1211
Relation	7803
Implied Intransitive Trajectory	1033

Fig 5.1: Total counts for each representation found in the Penn Treebank corpus

In the above chart, an ‘Of Possessive’ is a description that builds a possessive representation without using an apostrophe (e.g. “The penguin is the food of the bear”), whereas an ‘Apostrophe Possessive’ is its counterpart (e.g. “The penguin is the bear’s food”). The simple trajectory is just an object moving along a path (e.g. “The bird flew to the tree”), while the complex trajectory has an actor that moves the object along its path (e.g. “The boy threw the rock at the bird.”). The implied trajectories are either transitive, in which case they look like a Relation (e.g. “The woman climbed the mountain”) or they are intransitive (e.g. “The ship sank”).

It might be more interesting, however, to look at the data from the perspective of percentages of each representation.

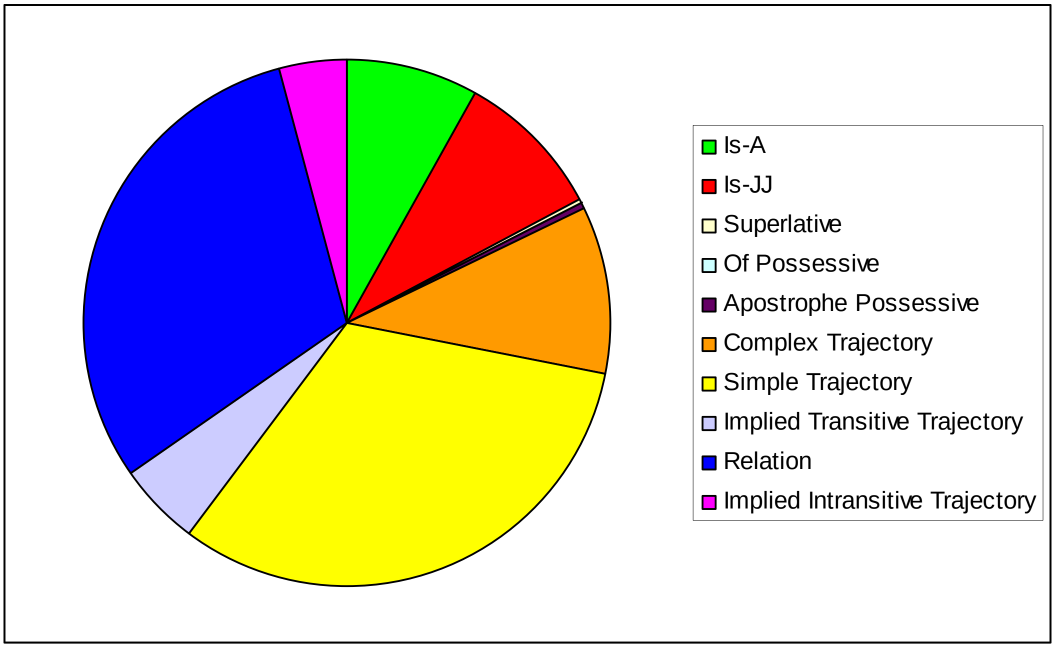

Fig 5.2: The relative presence of various representations in the Penn Treebank

5.3 Analysis

Of the representations used in the Penn Treebank, trajectories make up slightly over half. This was surprising to me, as I expected to see a lot more use of simple Is-A, Is-JJ, and a few more Relations. Of the trajectories, about 2/3 were simple go trajectories with prepositions, and of the remainder, a little more than half were trajectories where an actor sends an object to a target (like “The boy threw the ball to the girl”) and the rest was split half and half between implied transitive trajectories (like “The climber ascended the mountain”) and implied intransitive trajectories (like “The boat sank”). Relations make up about half of the other Penn Treebank representations, which is about as much as I expected. Finally, while superlatives and possessives were indeed rare, they were far rarer than I expected.

There were a few simple explanations for my mistaken assumptions, and they provided me with insights into the way the Wall Street Journal writers frame their statements to make them simpler and easier to digest for a reader.

First, superlatives were extremely rare in part because superlatives can be rewritten in a simpler and more compelling fashion as relations. For instance, “The US Dollar outperformed the Russian rubel” reads better, at least to me, than “The US Dollar is stronger than the Russian rubel”.

As for possessives, the low number is not surprising for two reasons. First and most importantly, declaring “X is the Y of the Z” isn't nearly as common of a concept as a relation (in fact, I sometimes think of the possessive representation as a subset of relations). Second, sentences that contain only a possessive are uncommon because the simple appositive structure is more common to express a possessive. So, for instance, instead of saying “Kenneth Lay is Enron's CEO.” and then “Lay broke several laws in his pursuit of wealth.”, a sentence might say “Kenneth Lay, Enron's CEO, broke several laws in his pursuit of wealth.” Because the task of rating the content of an appositive structure is too uncertain (sometimes the content of the interjectory phrase has no representation, and other times it might contain any sort with few cues), my system does not take any declarations made in appositives into account.

The lack of appositives is also a factor when it comes to Is-A and Is-JJ representations.

Additionally, Is-JJ can be avoided by a careful writer by using anaphora resolution and adjectives in another sort of representation. For instance, rather than saying “George W. Bush started a war in Iraq. George W. Bush is unpopular. George W. Bush is a president. George W. Bush has been criticised by members of both parties” an author might instead say “George W. Bush started a war in Iraq. The unpopular president has been criticised by members of both parties.” In this way, the author relies on the reader's ability to connect George W. Bush and president and thus encapsulates the Is-JJ relation implictly. In fact, the second set of sentences has also implied the Is-A relationship as well—the fact that George W. Bush is a president is implied by the fact that the reader must connect the 'president' in the second sentence with George W. Bush in the first.

The larger number of trajectories than I expected is undoubtably caused by metaphorical trajectories, which are more common in our language than it might seem. In the end, there were only simple go trajectories with a single object and its path in 1 of every 8 sentences on average, and that seems fairly reasonable, especially considering that some of the Penn Treebank sentences are quite long and contain several trajectories.

The coverage overall on the Penn Treebank was excellent. With just these few simple representations, the Span system managed to build useful descriptions at an average rate of one for every two sentences. When you consider the fact that I’m not even trying to handle verbs that require highly-specialised semantic knowledge to build descriptions (for instance, in the intransitive use of the verb ‘to drink’ such as in the sentence “He drank”, you need to know that there is an implied drink object or you cannot build a relation), that is an impressive amount of coverage.

Chapter 6

Authors and Representations

After examining the representations used in the Penn Treebank, I decided to apply the same analysis to several works of fiction in order to discover the techniques and representations used by various authors to build an imagined world in the reader's head.

6.1 Hypothesis

Before I began, I thought about what sorts of representations would make a good book. It would certainly vary according to the author's style, but the work should be evocative, with plenty of relations and trajectories and few boring Is-A sentences.

I chose eight works of literature: Oliver Twist and David Copperfield by Charles Dickens, Pride and Prejudice and Sense and Sensibility by Jane Austen, Heart of Darkness and Lord Jim by Joseph Conrad, and The Deerslayer and The Last of the Mohicans by James Fenimore Cooper. All of these works were available free in text format from Project Guteneberg.

Among these authors, there was a tendency towards long-winded sentences, particularly those with many semicolons, so I made the decision to cut sentences at end punctuation, semicolons, or colons in order to avoid sentences that ran hundreds of words long and broke the parser. I read in each sentence line by line and computed the representations.

6.2 Preliminary Results

The representations found in each work are as follows:

Pride and Prejudice	
Total Sentences	7624
Is-A	366
Is-JJ	838
Superlative	5
Of Possessive	6
Apostrophe Possessive	4
Complex Trajectory	229
Simple Trajectory	1104
Implied Transitive Trajectory	146
Relation	886
Implied Intransitive Trajectory	112

Sense and Sensibility	
Total Sentences	7078
Is-A	366
Is-JJ	820
Superlative	6
Of Possessive	5
Apostrophe Possessive	3
Complex Trajectory	247
Simple Trajectory	1318
Implied Transitive Trajectory	113
Relation	837
Implied Intransitive Trajectory	176

Fig 6.1 and 6.2: Total counts for each representation found in the novels *Pride and Prejudice* and *Sense and Sensibility*

The Deerslayer	
Total Sentences	9378
Is-A	421
Is-JJ	709
Superlative	10
Of Possessive	12
Apostrophe Possessive	7
Complex Trajectory	308
Simple Trajectory	1499
Implied Transitive Trajectory	159
Relation	1241
Implied Intransitive Trajectory	129

The Last of the Mohicans	
Total Sentences	7299
Is-A	236
Is-JJ	479
Superlative	12
Of Possessive	6
Apostrophe Possessive	0
Complex Trajectory	267
Simple Trajectory	1174
Implied Transitive Trajectory	118
Relation	722
Implied Intransitive Trajectory	83

Fig 6.3 and 6.4: Total counts for each representation found in the novels *The Deerslayer* and *The Last of the Mohicans*

Heart of Darkness	
Total Sentences	3413
Is-A	409
Is-JJ	658
Superlative	7
Of Possessive	12
Apostrophe Possessive	9
Complex Trajectory	262
Simple Trajectory	1401
Implied Transitive Trajectory	121
Relation	1003
Implied Intransitive Trajectory	115

Lord Jim	
Total Sentences	11130
Is-A	297
Is-JJ	527
Superlative	3
Of Possessive	7
Apostrophe Possessive	6
Complex Trajectory	218
Simple Trajectory	1318
Implied Transitive Trajectory	85
Relation	758
Implied Intransitive Trajectory	157

Fig 6.5 and 6.6: Total counts for each representation found in the novels *Heart of Darkness* and *Lord Jim*

David Copperfield	
Total Sentences	22560
Is-A	876
Is-JJ	1521
Superlative	17
Of Possessive	18
Apostrophe Possessive	16
Complex Trajectory	548
Simple Trajectory	3131
Implied Transitive Trajectory	286
Relation	2315
Implied Intransitive Trajectory	510

Oliver Twist	
Total Sentences	12581
Is-A	300
Is-JJ	347
Superlative	0
Of Possessive	6
Apostrophe Possessive	4
Complex Trajectory	271
Simple Trajectory	1264
Implied Transitive Trajectory	146
Relation	1127
Implied Intransitive Trajectory	155

Fig 6.7 and 6.8: Total counts for each representation found in the novels *David Copperfield* and *Oliver Twist*

The authors tended to use fewer representations per sentence overall. This is not surprising, however, considering that the novels contained significant dialogue, which often has filler sentences with no representations at all. Two of my favourites in this respect were the pithy exclamation “Ha!” and its counterpart “Ah!”.

The one effect I had not initially expected was the strong tendency towards Is-JJ representations in the novels compared to the Penn Treebank. This is likely a result of the difference in medium, and in hindsight I could have predicted it. The Penn Treebank sentences from the Wall Street Journal exist to report occurrences and thus do not need to focus on descriptions, whereas novels describe characters and their feelings in detail in order to paint a picture for the reader.

6.3 Comparing Authors and Styles

Using the data collected from the novels, I performed various statistical tests to explore three additional subgoals via three experiments:

1) Can I determine authorship by examining the similarities and differences in the use of representations—in other words, can I distinguish the author based on the way he or she imagines the scene?
2) If I compare the representations found in the novels to the Penn Treebank, are the two distributions distinct enough to be statistically significant?
3) If I instead pull out one novel and compare it to the rest, will it prove to be more similar than comparing the Penn Treebank to the novels?

Researchers have explored the problem of authorship attribution from multiple perspectives. Corey Gerritsen (Gerritsen 2003) experienced success with a system that determined a text's author by using Denis Yuret's theory of lexical attraction (Yuret 1999) and comparing the presence of words that were paired together. Thus far, however, comparisons of authorship have generally been based on statistical parsing methods involving the words themselves.

I decided to look at the problem from a new perspective, abstracted out a level from the text itself—can I find similarity between authors based purely on the way they use representations to help imagine their world? I call this kind of analysis 'representational analysis' to distinguish it from word-based and syntax-based textual analysis.

6.4 Book to Book Comparisons

I performed a t-test for each representation on each pair of books written by the same author. For the t-test, if the t-score is higher than t-critical, it means that the two samples (in this case the two books) were most likely drawn from two distinct sources. In each case, there is a probability listed which indicates the probability that both samples were taken from the same source and that thus the differences between them could be attributed to chance alone. These probabilities will be very low if the two sources were profoundly different. On the other hand, a low t-score means that the two samples are indistinguishable and might as well have been drawn from the same source. Two samples that are very similar will thus have a low t-score and a high probability that they were drawn from the same source.

t-Test: Austen and Is-A

	Variable 1	Variable 2
Mean	0.036201	0.036168
Variance	0.035683	0.036278
Observations	7624	7078
Hypothesized Mean Difference	0	
df	14600	
t Stat	0.010557	
P(T<=t) one-tail	0.495788	
t Critical one-tail	1.644958	
P(T<=t) two-tail	0.991577	
t Critical two-tail	1.960126	

t-Test: Austen and Is-JJ

	Variable 1	Variable 2
Mean	0.091159	0.087595
Variance	0.089682	0.090955
Observations	7624	7078
Hypothesized Mean Difference	0	
df	14603	
t Stat	0.718401	
P(T<=t) one-tail	0.236261	
t Critical one-tail	1.644958	
P(T<=t) two-tail	0.472522	
t Critical two-tail	1.960126	

Fig 6.9 and 6.10: T-tests for Is-A and Is-JJ representation use in *Pride and Prejudice* and *Sense and Sensibility*

t-Test: Austen and Superlative

	Variable 1	Variable 2
Mean	0.000393	0.000565
Variance	0.000393	0.000565
Observations	7624	7078
Hypothesized Mean Difference	0	
df	13822	
t Stat	-0.47348	
P(T<=t) one-tail	0.31794	
t Critical one-tail	1.644964	
P(T<=t) two-tail	0.63588	
t Critical two-tail	1.960136	

t-Test: Austen and Of Possessive

	Variable 1	Variable 2
Mean	0.000656	0.000283
Variance	0.000655	0.000283
Observations	7624	7078
Hypothesized Mean Difference	0	
df	13265	
t Stat	1.051987	
P(T<=t) one-tail	0.146412	
t Critical one-tail	1.644969	
P(T<=t) two-tail	0.292825	
t Critical two-tail	1.960143	

Fig 6.11 and 6.12: T-tests for Superlative and Of Possessive representation use in *Pride and Prejudice* and *Sense and Sensibility*

t-Test: Austen and Apostrophe Possessive

	Variable 1	Variable 2
Mean	0.000525	0.000283
Variance	0.000524	0.000283
Observations	7624	7078
Hypothesized Mean Difference	0	
df	13970	
t Stat	0.734271	
P(T<=t) one-tail	0.231398	
t Critical one-tail	1.644963	
P(T<=t) two-tail	0.462796	
t Critical two-tail	1.960134	

t-Test: Austen and Complex Trajectory

	Variable 1	Variable 2
Mean	0.022823	0.021899
Variance	0.022305	0.022835
Observations	7624	7078
Hypothesized Mean Difference	0	
df	14592	
t Stat	0.372466	
P(T<=t) one-tail	0.354776	
t Critical one-tail	1.644958	
P(T<=t) two-tail	0.709551	
t Critical two-tail	1.960127	

Fig 6.13 and 6.14: T-tests for Apostrophe Possessive and Complex Trajectory representation use in *Pride and Prejudice* and *Sense and Sensibility*

t-Test: Austen and Simple Trajectory

	Variable 1	Variable 2
Mean	0.110178	0.113733
Variance	0.114843	0.118899
Observations	7624	7078
Hypothesized Mean Difference	0	
df	14578	
t Stat	-0.62968	
P(T<=t) one-tail	0.264457	
t Critical one-tail	1.644958	
P(T<=t) two-tail	0.528913	
t Critical two-tail	1.960127	

t-Test: Austen and Implied Transitive Trajectory

	Variable 1	Variable 2
Mean	0.015477	0.011726
Variance	0.01524	0.011591
Observations	7624	7078
Hypothesized Mean Difference	0	
df	14643	
t Stat	1.966988	
P(T<=t) one-tail	0.024602	
t Critical one-tail	1.644958	
P(T<=t) two-tail	0.049203	
t Critical two-tail	1.960126	

Fig 6.15 and 6.16: T-tests for Simple Trajectory and Implied Transitive Trajectory representation use in *Pride and Prejudice* and *Sense and Sensibility*

t-Test: Austen and Implied Intransitive Trajectory

	Variable 1	Variable 2
Mean	0.011674	0.014552
Variance	0.011539	0.014342
Observations	7624	7078
Hypothesized Mean Difference	0	
df	14228	
t Stat	-1.52993	
P(T<=t) one-tail	0.063029	
t Critical one-tail	1.644961	
P(T<=t) two-tail	0.126057	
t Critical two-tail	1.960131	

t-Test: Austen and Relation

	Variable 1	Variable 2
Mean	0.088536	0.077282
Variance	0.09645	0.079515
Observations	7624	7078
Hypothesized Mean Difference	0	
df	14693	
t Stat	2.302839	
P(T<=t) one-tail	0.010651	
t Critical one-tail	1.644957	
P(T<=t) two-tail	0.021302	
t Critical two-tail	1.960125	

Fig 6.17 and 6.18: T-tests for Implies Intransitive Trajectory and Relation representation use in *Pride and Prejudice* and *Sense and Sensibility*

t-Test: Conrad and Is-A

	Variable 1	Variable 2
Mean	0.034281	0.026685
Variance	0.03546	0.027772
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5157	
t Stat	2.116156	
P(T<=t) one-tail	0.01719	
t Critical one-tail	1.645149	
P(T<=t) two-tail	0.034379	
t Critical two-tail	1.960424	

t-Test: Conrad and Is-JJ

	Variable 1	Variable 2
Mean	0.050396	0.04735
Variance	0.052559	0.048346
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5475	
t Stat	0.685526	
P(T<=t) one-tail	0.246521	
t Critical one-tail	1.645132	
P(T<=t) two-tail	0.493041	
t Critical two-tail	1.960397	

Fig 6.19 and 6.20: T-tests for Is-A and Is-JJ representation use in *Heart of Darkness* and *Lord Jim*

t-Test: Conrad and Superlative

	Variable 1	Variable 2
Mean	0.000293	0.00027
Variance	0.000293	0.000269
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5475	
t Stat	0.070702	
P(T<=t) one-tail	0.471819	
t Critical one-tail	1.645132	
P(T<=t) two-tail	0.943638	
t Critical two-tail	1.960397	

t-Test: Conrad and Of Possessive

	Variable 1	Variable 2
Mean	0.000879	0.000629
Variance	0.000878	0.000629
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5000	
t Stat	0.446347	
P(T<=t) one-tail	0.327683	
t Critical one-tail	1.645158	
P(T<=t) two-tail	0.655366	
t Critical two-tail	1.960438	

Fig 6.21 and 6.22: T-tests for Superlative and Of Possessive representation use in *Heart of Darkness* and *Lord Jim*

t-Test: Conrad and Apostrophe Possessive

	Variable 1	Variable 2
Mean	0.000586	0.000539
Variance	0.000586	0.000539
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5475	
t Stat	0.100002	
P(T<=t) one-tail	0.460173	
t Critical one-tail	1.645132	
P(T<=t) two-tail	0.920347	
t Critical two-tail	1.960397	

t-Test: Conrad and Complex Trajectory

	Variable 1	Variable 2
Mean	0.019045	0.019587
Variance	0.018688	0.019384
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5749	
t Stat	-0.20171	
P(T<=t) one-tail	0.420077	
t Critical one-tail	1.645119	
P(T<=t) two-tail	0.840154	
t Critical two-tail	1.960377	

Fig 6.23 and 6.24: T-tests for Apostrophe Possessive and Complex Trajectory representation use in *Heart of Darkness* and *Lord Jim*

t-Test: Conrad and Simple Trajectory

	Variable 1	Variable 2
Mean	0.133021	0.118419
Variance	0.134704	0.121118
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5426	
t Stat	2.057862	
P(T<=t) one-tail	0.019825	
t Critical one-tail	1.645135	
P(T<=t) two-tail	0.039651	
t Critical two-tail	1.960401	

t-Test: Conrad and Implied Transitive Trajectory

	Variable 1	Variable 2
Mean	0.007325	0.007637
Variance	0.007273	0.007579
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5761	
t Stat	-0.1861	
P(T<=t) one-tail	0.426185	
t Critical one-tail	1.645118	
P(T<=t) two-tail	0.852369	
t Critical two-tail	1.960376	

Fig 6.25 and 6.26: T-tests for Simple Trajectory and Implied Transitive Trajectory representation use in *Heart of Darkness* and *Lord Jim*

t-Test: Conrad and Relation

	Variable 1	Variable 2
Mean	0.073835	0.068104
Variance	0.076024	0.072278
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5547	
t Stat	1.06851	
P(T<=t) one-tail	0.142669	
t Critical one-tail	1.645128	
P(T<=t) two-tail	0.285337	
t Critical two-tail	1.960392	

t-Test: Conrad and Implied Intransitive Trajectory

	Variable 1	Variable 2
Mean	0.012306	0.014106
Variance	0.012158	0.013908
Observations	3413	11130
Hypothesized Mean Difference	0	
df	5999	
t Stat	-0.82063	
P(T<=t) one-tail	0.205945	
t Critical one-tail	1.645108	
P(T<=t) two-tail	0.411891	
t Critical two-tail	1.960359	

Fig 6.27 and 6.28: T-tests for Relation and Implied Intransitive Trajectory representation use in *Heart of Darkness* and *Lord Jim*

Taking a look at the data, we can see that both Joseph Conrad and Jane Austen are extremely consistent in their use of representations between the two books I sampled for each—in fact, for many of the representations, the two books for each of these authors were considered indistinguishable with up to 90% probability. The few representations with significant differences were only just barely so, with probabilities like 3% or 2% that the two sources might have been the same and the changes in representation style were due to chance. In general, 5% is seen as a cutoff in the world of statistics—any higher probability and it might actually just be by chance. Since those differences are only barely below the cutoff (and since we'll see later that extremely different texts produce results that are astronomically lower!), they may have simply been due to chance as well.

t-Test: Cooper and Is-A

	Variable 1	Variable 2
Mean	0.044892	0.032333
Variance	0.046934	0.032388
Observations	9378	7299
Hypothesized Mean Difference	0	
df	16605	
t Stat	4.087197	
P(T<=t) one-tail	2.19E-05	
t Critical one-tail	1.644945	
P(T<=t) two-tail	4.39E-05	
t Critical two-tail	1.960107	

t-Test: Cooper and Is-JJ

	Variable 1	Variable 2
Mean	0.075602	0.065625
Variance	0.083118	0.067904
Observations	9378	7299
Hypothesized Mean Difference	0	
df	16308	
t Stat	2.340823	
P(T<=t) one-tail	0.009627	
t Critical one-tail	1.644947	
P(T<=t) two-tail	0.019253	
t Critical two-tail	1.960109	

Fig 6.29 and 6.30: T-tests for Is-A and Is-JJ representation use in *The Deerslayer* and *The Last of the Mohicans*

t-Test: Cooper and Superlative

	Variable 1	*Variable 2*
Mean	0.001066	0.001644
Variance	0.001065	0.001642
Observations	9378	7299
Hypothesized Mean Difference	0	
df	13793	
t Stat	-0.993	
P(T<=t) one-tail	0.160364	
t Critical one-tail	1.644964	
P(T<=t) two-tail	0.320728	
t Critical two-tail	1.960136	

t-Test: Cooper and Of Possessive

	Variable 1	*Variable 2*
Mean	0.00128	0.000822
Variance	0.001278	0.000821
Observations	9378	7299
Hypothesized Mean Difference	0	
df	16661	
t Stat	0.917267	
P(T<=t) one-tail	0.179508	
t Critical one-tail	1.644945	
P(T<=t) two-tail	0.359016	
t Critical two-tail	1.960106	

Fig 6.31 and 6.32: T-tests for Superlative and Of Possessive representation use in *The Deerslayer* and *The Last of the Mohicans*

t-Test: Cooper and Apostrophe Possessive

	Variable 1	*Variable 2*
Mean	0.000746	0
Variance	0.000746	0
Observations	9378	7299
Hypothesized Mean Difference	0	
df	9377	
t Stat	2.646598	
P(T<=t) one-tail	0.004072	
t Critical one-tail	1.645016	
P(T<=t) two-tail	0.008144	
t Critical two-tail	1.960217	

t-Test: Cooper and Complex Trajectory

	Variable 1	*Variable 2*
Mean	0.032843	0.03658
Variance	0.033474	0.037165
Observations	9378	7299
Hypothesized Mean Difference	0	
df	15274	
t Stat	-1.26997	
P(T<=t) one-tail	0.102057	
t Critical one-tail	1.644953	
P(T<=t) two-tail	0.204114	
t Critical two-tail	1.960119	

Fig 6.33 and 6.34: T-tests for Apostrophe Possessive and Complex Trajectory representation use in *The Deerslayer* and *The Last of the Mohicans*

t-Test: Cooper and Simple Trajectory

	Variable 1	*Variable 2*
Mean	0.159842	0.160844
Variance	0.159048	0.1613
Observations	9378	7299
Hypothesized Mean Difference	0	
df	15632	
t Stat	-0.16029	
P(T<=t) one-tail	0.436327	
t Critical one-tail	1.644951	
P(T<=t) two-tail	0.872654	
t Critical two-tail	1.960116	

t-Test: Cooper and Implied Transitive Trajectory

	Variable 1	*Variable 2*
Mean	0.016955	0.016167
Variance	0.016669	0.015907
Observations	9378	7299
Hypothesized Mean Difference	0	
df	15851	
t Stat	0.396131	
P(T<=t) one-tail	0.346007	
t Critical one-tail	1.64495	
P(T<=t) two-tail	0.692014	
t Critical two-tail	1.960114	

Fig 6.35 and 6.36: T-tests for Simple Trajectory and Implied Transitive Trajectory representation use in *The Deerslayer* and *The Last of the Mohicans*

t-Test: Cooper and Relation

	Variable 1	Variable 2
Mean	0.132331	0.098918
Variance	0.139147	0.104492
Observations	9378	7299
Hypothesized Mean Difference	0	
df	16484	
t Stat	6.188347	
P(T<=t) one-tail	3.11E-10	
t Critical one-tail	1.644946	
P(T<=t) two-tail	6.22E-10	
t Critical two-tail	1.960108	

t-Test: Cooper and Implied Intransitive Trajectory

	Variable 1	Variable 2
Mean	0.013756	0.011371
Variance	0.013568	0.011244
Observations	9378	7299
Hypothesized Mean Difference	0	
df	16273	
t Stat	1.379449	
P(T<=t) one-tail	0.083888	
t Critical one-tail	1.644947	
P(T<=t) two-tail	0.167775	
t Critical two-tail	1.96011	

Fig 6.37 and 6.38: T-tests for Relation and Implied Intransitive Trajectory representation use in *The Deerslayer* and *The Last of the Mohicans*

James Fenimore Cooper's two works were generally similar, but *The Deerslayer* had significantly more Is-A and Relation representations than *The Last of the Mohicans*. Because *The Deerslayer* was the first in the *Leatherstocking Tales* series and *The Last of the Mohicans* is second, it makes sense there needed to be more Is-A representations to set the scene and explain the way Cooper's world works. The difference in relations is interesting—perhaps Cooper tended to show the relations in *The Last of the Mohicans* through actions and dialogue, rather than stating them outright as he did in *The Deerslayer*. This might be an indication of Cooper improving his descriptive technique between the two books.

t-Test: Dickens and Is-A

	Variable 1	Variable 2
Mean	0.03883	0.023845
Variance	0.039451	0.024074
Observations	22560	12581
Hypothesized Mean Difference	0	
df	31438	
t Stat	7.830042	
P(T<=t) one-tail	2.51E-15	
t Critical one-tail	1.644902	
P(T<=t) two-tail	5.03E-15	
t Critical two-tail	1.960039	

t-Test: Dickens and Is-JJ

	Variable 1	Variable 2
Mean	0.06742	0.027581
Variance	0.071832	0.028571
Observations	22560	12581
Hypothesized Mean Difference	0	
df	34627	
t Stat	17.05725	
P(T<=t) one-tail	2.85E-65	
t Critical one-tail	1.644898	
P(T<=t) two-tail	5.69E-65	
t Critical two-tail	1.960032	

Fig 6.39 and 6.40: T-tests for Is-A and Is-JJ representation use in *David Copperfield* and *Oliver Twist*

t-Test: Dickens and Superlative

	Variable 1	*Variable 2*
Mean	0.000754	0
Variance	0.000753	0
Observations	22560	12581
Hypothesized Mean Difference	0	
df	22559	
t Stat	4.124569	
P(T<=t) one-tail	1.86E-05	
t Critical one-tail	1.644921	
P(T<=t) two-tail	3.73E-05	
t Critical two-tail	1.960069	

t-Test: Dickens and Of Possessive

	Variable 1	*Variable 2*
Mean	0.000798	0.000477
Variance	0.000797	0.000477
Observations	22560	12581
Hypothesized Mean Difference	0	
df	31640	
t Stat	1.186052	
P(T<=t) one-tail	0.117805	
t Critical one-tail	1.644902	
P(T<=t) two-tail	0.235611	
t Critical two-tail	1.960039	

Fig 6.41 and 6.42: T-tests for Superlative and Of Possessive representation use in *David Copperfield* and *Oliver Twist*

t-Test: Dickens and Apostrophe Possessive

	Variable 1	*Variable 2*
Mean	0.000709	0.000318
Variance	0.000709	0.000318
Observations	22560	12581
Hypothesized Mean Difference	0	
df	34000	
t Stat	1.64349	
P(T<=t) one-tail	0.050145	
t Critical one-tail	1.644898	
P(T<=t) two-tail	0.100291	
t Critical two-tail	1.960034	

t-Test: Dickens and Complex Trajectory

	Variable 1	*Variable 2*
Mean	0.024291	0.02154
Variance	0.024145	0.021555
Observations	22560	12581
Hypothesized Mean Difference	0	
df	27271	
t Stat	1.648502	
P(T<=t) one-tail	0.049631	
t Critical one-tail	1.64491	
P(T<=t) two-tail	0.099261	
t Critical two-tail	1.960051	

Fig 6.43 and 6.44: T-tests for Apostrophe Possessive and Complex Trajectory representation use in *David Copperfield* and *Oliver Twist*

t-Test: Dickens and Simple Trajectory

	Variable 1	*Variable 2*
Mean	0.138785	0.100469
Variance	0.141871	0.103101
Observations	22560	12581
Hypothesized Mean Difference	0	
df	29581	
t Stat	10.06812	
P(T<=t) one-tail	4.17E-24	
t Critical one-tail	1.644905	
P(T<=t) two-tail	8.35E-24	
t Critical two-tail	1.960044	

t-Test: Dickens and Implied Transitive Trajectory

	Variable 1	*Variable 2*
Mean	0.012677	0.011605
Variance	0.012694	0.011471
Observations	22560	12581
Hypothesized Mean Difference	0	
df	27135	
t Stat	0.883244	
P(T<=t) one-tail	0.188556	
t Critical one-tail	1.64491	
P(T<=t) two-tail	0.377112	
t Critical two-tail	1.960051	

Fig 6.45 and 6.46: T-tests for Simple Trajectory and Implied Transitive Trajectory representation use in *David Copperfield* and *Oliver Twist*

t-Test: Dickens and Relation

	Variable 1	*Variable 2*
Mean	0.102615	0.08958
Variance	0.108402	0.088557
Observations	22560	12581
Hypothesized Mean Difference	0	
df	28271	
t Stat	3.787792	
P(T<=t) one-tail	7.62E-05	
t Critical one-tail	1.644908	
P(T<=t) two-tail	0.000152	
t Critical two-tail	1.960048	

t-Test: Dickens and Implied Intransitive Trajectory

	Variable 1	*Variable 2*
Mean	0.022606	0.01232
Variance	0.022096	0.012169
Observations	22560	12581
Hypothesized Mean Difference	0	
df	32419	
t Stat	7.372305	
P(T<=t) one-tail	8.59E-14	
t Critical one-tail	1.644901	
P(T<=t) two-tail	1.72E-13	
t Critical two-tail	1.960037	

Fig 6.47 and 6.48: T-tests for Relation and Implied Intransitive Trajectory representation use in *David Copperfield* and *Oliver Twist*

Charles Dickens turned out to be the black sheep of the crop, as his works were vastly divergent in nearly every representation. This shows that Dickens is a versatile author, able to change the way he describes and imagines his world based on the genre and motif of the work.

Oliver Twist was a social novel (a novel that calls a social ill to attention), the first in the English language to focus throughout on a child protagonist. It was only Dickens's second novel, published in 1838.

On the other hand, *David Copperfield* was a bildungsroman (a novel of personal development and maturity) with autobiographical elements, and it was written by a more mature Dickens in 1850.

These differences led Dickens to use a different style of representation in each of the two works. In constrast, Lord Jim and Heart of Darkness are both stories of the adventures of Marlow, The Deerslayer and The Last of the Mohicans are both adventures of Natti Bumppo in the same series, and Pride and Prejudice and Sense and Sensibility are both romances in Jane Austen's signature style. Thus, Oliver Twist and David Copperfield were the most different to start.

Comparing books across authors leads to results similar to that of the two Dickens novels—The representations from the two works are found to be different with an extremely low probability that they might be from the same distribution. There are two main exceptions, however. Often the possessives or the superlatives will appear to be the same across texts with relatively high probability. This occurs because they are so rare, and their use seems fairly standard in that rarity throughout various texts. Also, sometimes the implied trajectories will appear to be the same across the two works. This is more a case of the usage of particular verbs because implied trajectories are located by the use of the trajectory verbs. Because all of the works were from roughly the same time period, it makes sense that the use of verbs might be similar occasionally between two texts.

6.5 The Penn Treebank and the Novels

Next I compared the Penn Treebank sentences with the amalgam of all the novels, using the same t-test to determine whether the Penn Treebank and the novels were significantly different. As before, a t-score higher than t-critical means that the two samples were definitely different, whereas a low t-score means that they were nearly indistinguishable.

t-Test: Penn vs Authors for Is-A

	Variable 1	Variable 2
Mean	0.041394	0.036286
Variance	0.040902	0.037048
Observations	49162	65425
Hypothesized Mean Difference	0	
df	103002	
t Stat	4.319695	
P(T<=t) one-tail	7.82E-06	
t Critical one-tail	1.644868	
P(T<=t) two-tail	1.56E-05	
t Critical two-tail	1.959987	

t-Test: Penn vs Authors for Is-JJ

	Variable 1	Variable 2
Mean	0.047557	0.069362
Variance	0.048551	0.072439
Observations	49162	65425
Hypothesized Mean Difference	0	
df	113750	
t Stat	-15.0655	
P(T<=t) one-tail	1.53E-51	
t Critical one-tail	1.644867	
P(T<=t) two-tail	3.06E-51	
t Critical two-tail	1.959985	

Fig 6.49 and 6.50: T-tests for Is-A and Is-JJ representation use in the Penn Treebank and the novels

t-Test: Penn vs Authors for Superlative

	Variable 1	Variable 2
Mean	0.000976	0.000718
Variance	0.000975	0.000718
Observations	49162	65425
Hypothesized Mean Difference	0	
df	96413	
t Stat	1.469675	
P(T<=t) one-tail	0.070826	
t Critical one-tail	1.644869	
P(T<=t) two-tail	0.141653	
t Critical two-tail	1.959989	

t-Test: Penn vs Authors for Of Possessive

	Variable 1	Variable 2
Mean	0.000895	0.000749
Variance	0.000894	0.000748
Observations	49162	65425
Hypothesized Mean Difference	0	
df	100554	
t Stat	0.8485	
P(T<=t) one-tail	0.198081	
t Critical one-tail	1.644869	
P(T<=t) two-tail	0.396162	
t Critical two-tail	1.959988	

Fig 6.51 and 6.52: T-tests for Superlative and Of Possessive representation use in the Penn Treebank and the novels

t-Test: Penn vs Authors for Apostrophe Possessive

	Variable 1	Variable 2
Mean	0.000427	0.00052
Variance	0.000427	0.000519
Observations	49162	65425
Hypothesized Mean Difference	0	
df	110644	
t Stat	-0.71757	
P(T<=t) one-tail	0.236513	
t Critical one-tail	1.644867	
P(T<=t) two-tail	0.473026	
t Critical two-tail	1.959985	

t-Test: Penn vs Authors for Complex Trajectory

	Variable 1	Variable 2
Mean	0.053619	0.025541
Variance	0.058474	0.025653
Observations	49162	65425
Hypothesized Mean Difference	0	
df	80354	
t Stat	22.32688	
P(T<=t) one-tail	2.2E-110	
t Critical one-tail	1.644873	
P(T<=t) two-tail	4.4E-110	
t Critical two-tail	1.959993	

Fig 6.53 and 6.54: T-tests for Apostrophe Possessive and Complex Trajectory representation use in the Penn Treebank and the novels

t-Test: Penn vs Authors for Simple Trajectory

	Variable 1	*Variable 2*
Mean	0.166938	0.134872
Variance	0.180203	0.137899
Observations	49162	65425
Hypothesized Mean Difference	0	
df	97683	
t Stat	13.34548	
P(T<=t) one-tail	6.83E-41	
t Critical one-tail	1.644869	
P(T<=t) two-tail	1.37E-40	
t Critical two-tail	1.959988	

t-Test: Penn/Authors for Implied Transitive Trajectory

	Variable 1	*Variable 2*
Mean	0.024633	0.012717
Variance	0.024027	0.012616
Observations	49162	65425
Hypothesized Mean Difference	0	
df	85597	
t Stat	14.43371	
P(T<=t) one-tail	1.8E-47	
t Critical one-tail	1.644871	
P(T<=t) two-tail	3.61E-47	
t Critical two-tail	1.959992	

Fig 6.55 and 6.56: T-tests for Simple Trajectory and Implied Transitive Trajectory representation use in the Penn Treebank and the novels

t-Test: Penn vs Authors for Relation

	Variable 1	*Variable 2*
Mean	0.15872	0.094948
Variance	0.165955	0.100761
Observations	49162	65425
Hypothesized Mean Difference	0	
df	90151	
t Stat	28.76288	
P(T<=t) one-tail	2.1E-181	
t Critical one-tail	1.644871	
P(T<=t) two-tail	4.1E-181	
t Critical two-tail	1.95999	

t-Test: Penn/Authors for Implied Intransitive Trajectory

	Variable 1	*Variable 2*
Mean	0.021012	0.015621
Variance	0.020571	0.015377
Observations	49162	65425
Hypothesized Mean Difference	0	
df	96921	
t Stat	6.669216	
P(T<=t) one-tail	1.29E-11	
t Critical one-tail	1.644869	
P(T<=t) two-tail	2.59E-11	
t Critical two-tail	1.959988	

Fig 6.57 and 6.58: T-tests for Relation and Implied Intransitive Trajectory representation use in the Penn Treebank and the novels

As before between different authors, the superlatives and the possessives are rare and used in about equal proportions, such that their use in the Penn Treebank was not distinguishable from their use in the novels. However, for every other representation, the Penn Treebank was significantly different from the novels, often with extremely high probability (the probability that the difference in the use of Relations between the novels and the Penn Treebank is just due to chance is less likely than the probability of choosing a random atom in the universe three times and picking the same atom all three times!).

6.6 One Novel Among Many

Finally, I compared David Copperfield to all of the rest of the novels in an amalgam, just to make sure that the amalgam was not what caused the incredibly strong results with the Penn Treebank.

t-Test: DC vs Others for Is-A

	Variable 1	*Variable 2*
Mean	0.03883	0.034905
Variance	0.039451	0.035648
Observations	22560	45924
Hypothesized Mean Difference	0	
df	42881	
t Stat	2.469639	
P(T<=t) one-tail	0.006764	
t Critical one-tail	1.644889	
P(T<=t) two-tail	0.013529	
t Critical two-tail	1.960019	

t-Test: DC vs Others for Is-JJ

	Variable 1	*Variable 2*
Mean	0.06742	0.069724
Variance	0.071832	0.072572
Observations	22560	45924
Hypothesized Mean Difference	0	
df	45056	
t Stat	-1.05541	
P(T<=t) one-tail	0.145621	
t Critical one-tail	1.644887	
P(T<=t) two-tail	0.291243	
t Critical two-tail	1.960017	

Fig 6.59 and 6.60: T-tests for Is-A and Is-JJ representation use in David Copperfield and the other novels

t-Test: DC vs Others for Superlative

	Variable 1	*Variable 2*
Mean	0.000754	0.000719
Variance	0.000753	0.000718
Observations	22560	45924
Hypothesized Mean Difference	0	
df	43912	
t Stat	0.157944	
P(T<=t) one-tail	0.437251	
t Critical one-tail	1.644888	
P(T<=t) two-tail	0.874501	
t Critical two-tail	1.960018	

t-Test: DC vs Others for Of Possessive

	Variable 1	*Variable 2*
Mean	0.000798	0.000762
Variance	0.000797	0.000762
Observations	22560	45924
Hypothesized Mean Difference	0	
df	43944	
t Stat	0.156862	
P(T<=t) one-tail	0.437677	
t Critical one-tail	1.644888	
P(T<=t) two-tail	0.875354	
t Critical two-tail	1.960018	

Fig 6.61 and 6.62: T-tests for Superlative and Of Possessive representation use in David Copperfield and the other novels

t-Test: DC vs Others for Apostrophe Possessive

	Variable 1	*Variable 2*
Mean	0.000709	0.000457
Variance	0.000709	0.000457
Observations	22560	45924
Hypothesized Mean Difference	0	
df	37279	
t Stat	1.238693	
P(T<=t) one-tail	0.107733	
t Critical one-tail	1.644895	
P(T<=t) two-tail	0.215467	
t Critical two-tail	1.960028	

t-Test: DC vs Others for Complex Trajectory

	Variable 1	*Variable 2*
Mean	0.024291	0.025847
Variance	0.024145	0.026094
Observations	22560	45924
Hypothesized Mean Difference	0	
df	46441	
t Stat	-1.21581	
P(T<=t) one-tail	0.112031	
t Critical one-tail	1.644886	
P(T<=t) two-tail	0.224062	
t Critical two-tail	1.960015	

Fig 6.63 and 6.64: T-tests for Apostrophe Possessive and Complex Trajectory representation use in David Copperfield and the other novels

t-Test: DC vs Others for Simple Trajectory

	Variable 1	*Variable 2*
Mean	0.138785	0.132632
Variance	0.141871	0.135338
Observations	22560	45924
Hypothesized Mean Difference	0	
df	43919	
t Stat	2.024764	
P(T<=t) one-tail	0.021449	
t Critical one-tail	1.644888	
P(T<=t) two-tail	0.042898	
t Critical two-tail	1.960018	

t-Test: DC/Others for Implied Transitive Trajectory

	Variable 1	*Variable 2*
Mean	0.012677	0.012804
Variance	0.012694	0.01264
Observations	22560	45924
Hypothesized Mean Difference	0	
df	44764	
t Stat	-0.13815	
P(T<=t) one-tail	0.445063	
t Critical one-tail	1.644888	
P(T<=t) two-tail	0.890125	
t Critical two-tail	1.960017	

Fig 6.65 and 6.66: T-tests for Simple Trajectory and Implied Transitive Trajectory representation use in David Copperfield and the other novels

t-Test: DC vs Others for Relation

	Variable 1	*Variable 2*
Mean	0.102615	0.091368
Variance	0.108402	0.097002
Observations	22560	45924
Hypothesized Mean Difference	0	
df	42698	
t Stat	4.276268	
P(T<=t) one-tail	9.52E-06	
t Critical one-tail	1.644889	
P(T<=t) two-tail	1.9E-05	
t Critical two-tail	1.960019	

t-Test: DC/Others for Implied Intransitive Trajectory

	Variable 1	*Variable 2*
Mean	0.022606	0.01313
Variance	0.022096	0.012958
Observations	22560	45924
Hypothesized Mean Difference	0	
df	35963	
t Stat	8.436484	
P(T<=t) one-tail	1.7E-17	
t Critical one-tail	1.644896	
P(T<=t) two-tail	3.39E-17	
t Critical two-tail	1.96003	

Fig 6.67 and 6.69: T-tests for Relation and Implied Intransitive Trajectory representation use in David Copperfield and the other novels

David Copperfield was quite different from the amalgam when it came to the use of Relations (which is not surprising, considering that David Copperfield was abnormal in that regard even in comparison to Oliver Twist in the earlier study). It was also quite different in the use of intransitive implied trajectories, which just means that Dickens used those particular verbs more than the other authors in the amalgam (because we know already that David Copperfield and Oliver Twist were similar in that regard). There may be slight differences in the use of Is-A and simple trajectories, but the effect is not very strong considering the number of samples, so it may just be noise. Otherwise, David Copperfield proved to be passingly similar to the amalgam of the other novels, which just shows how striking the difference is between the amalgam and the Penn Treebank.

The three experiments I performed show that studying similarities and differences in texts and authors on a representational level, rather than a purely textual level, can produce nuanced and thought-provoking results. I hope that others will use the Span system and representational analysis further in the future to extend these findings to other applications.

Chapter 7

Contributions

Throughout this thesis, I have explored the question of what we can learn about how we think from what we say. Specifically, I have:

- Motivated the study of the role of language in thought, as opposed to purely logic-based approaches, as crucial to the success of Artificial Intelligence
- Tied together the work of Artificial Intelligence researchers with a broader field-spanning investigation into the connections between human language and human thought
- Created a system that can build cognitive representations based on a natural language input. For example, in the sentence "The dog walked across the street", the system would use the embedded description of movement along a path to instantiate a general purpose trajectory representation that models movement along a path
- Illustrated a way in which my system can transform a descriptive sentence into a three-dimensional graphical scene. For instance, the sentence "The bird flew to the top of the tree" allows us to imagine a three-dimensional scene wherein a bird model flies up to the top of a tree model (see pages 14-17 for some pictures).
- Produced a detailed analysis of the types of representations used in the Penn Treebank from the Wall Street Journal corpus. For instance, my system finds that the Penn Treebank has one embedded description for a trajectory representation in around every four sentences, on average.
- Developed the concept of representational analysis, an analysis of texts that focuses on the representational level rather than the surface level
- Used representational analysis to explore the styles of four major authors and examined the way they used representations to tell their stories and stimulate our imaginations
- Discovered representational similarities and differences among those authors—for instance:
  - Dickens is more versatile than the others in the way he describes his world—his use of representations varies greatly from novel to novel . For example, the difference between *David Copperfield* and *Oliver Twist* in their use of representations was much larger than that between two works of other authors.
  - Jane Austen is very consistent, and her novels are indistinguishable from each other (and thus clearly in her own style) from the viewpoint of representational analysis. In one extreme case, her use of Is-A representations between *Pride and Prejudice* and *Sense and Sensibility* was almost completely identical!
  - When combined into a single dataset, the descriptions in the novels and the representations they used were vastly different than those in the Penn Treebank. For instance, the Penn Treebank placed more emphasis on Is-A representations (e.g. "Pierre Vinken is the CEO") and significantly less emphasis on descriptive Is-JJ representations. (e.g. "Mrs Dashwood was surprised")

References:

Austen, J. (1813). *Pride and Prejudice.*Via Project Gutenberg http://www.gutenberg.org/wiki/Main_Page

Austen, J. (1811). *Sense and Sensibility*. Via Project Gutenberg http://www.gutenberg.org/wiki/Main_Page

Bender, J. R. (2001). *Connecting Language and Vision Using a Conceptual Semantics.* Masters of Engineering Thesis, MIT. Cambridge, MA.

Bonawitz, K. (2003). *Bidirectional Natural Language Parsing using Streams and Counterstreams.* Masters of Engineering Thesis, MIT. Cambridge, MA.

Borchardt, G. C. (1994). *Thinking between the Lines: Computers and the Comprehension of Causal Descriptions.* Cambridge, MA, MIT Press.

Conrad, J. (1899). *Heart of Darkness*. Via Project Gutenberg http://www.gutenberg.org/wiki/Main_Page

Conrad, J. (1900). *Lord Jim*. Via Project Gutenberg http://www.gutenberg.org/wiki/Main_Page

Cooper, H. Personal conversations.

Cooper, J. (1841).*The Deerslayer*. Via Project Gutenberg http://www.gutenberg.org/wiki/Main_Page

Cooper, J. (1854). *The Last of the Mohicans*. Via Project Gutenberg http://www.gutenberg.org/wiki/Main_Page

Dickens, C. (1850). *David Copperfield.* Via Project Gutenberg http://www.gutenberg.org/wiki/Main_Page

Dickens, C. (1838).*Oliver Twist.* Via Project Gutenberg http://www.gutenberg.org/wiki/Main_Page

Gerritsen, C. (2003). *Authorship Attribution Using Lexical Attraction*. Masters of Engineering Thesis, MIT. Cambridge, MA.

Greenblatt, R. and L. Vaina (1979). "The Use of Thread Memory in Amnesic Aphasia and Concept Learning." AI Working Paper 195, MIT. Cambridge, MA.

Hearst, M. (1992). *Automatic Acquisition of Hyponyms from Large Text Corpora*. In Proceedings of the Fourteenth International Conference on Computational Linguistics. Nantes, France.

Jackendoff, R. (1983). *Semantics and cognition.* Cambridge, MA. MIT Press.

Katz, Boris. (1997). *An Overview of the START system*. In Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97).

Larson, S. (2003). *Intrinsic Representation: Boostrapping Symbols from Experience.* Masters of Engineering Thesis, MIT. Cambridge, MA.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*.

Manning, C.D. (2007). Stanford Natural Language Processing Group Parser, http://nlp.stanford.edu/downloads/lex-parser.shtml

Moore, D. Personal conversations.
Turing, A. (1963). "Computing Machinery and Intelligence." *Computers and Thought*: 11-35

Orwell, G. (1948). *1984*.

Sapir, E. (1929). "The Status of Linguistics as a Science." *Language* 5:209.

Shadadi, A. (2003). *Barnyard Politics: A Decision Rationale Representation for the Analysis of Simple Political Situations.* Masters of Engineering Thesis, MIT. Cambridge, MA.

Spelke, E. , Hermer-Vasquez, L. , and A. Katsnelson. (1999). "Sources of Flexibility in Human Cognition: Dual-Task Studies of Space and Language." *Cognitive Psychology 39*: 3-36

Turing, A. (1963). "Computing Machiney and Intelligence". *Computers and Thought*: 11-35

Tversky, Amos, and Daniel Kahneman (1981). "The Framing of Decisions and the Psychology of Choice." Science 211: 453-458.

Whorf, B. (1956). *Language, Thought & Reality*. Cambridge, MA: MIT Press.

Wittgenstein, L. (1966). *Tractatus Logico-Philosophicus.*

Yuret, D. (1999). *Lexical Attraction Models of Language.* In Proceedings of AAAI 1999.