

TRACKING THE EVOLUTION OF SCIENCE

A THESIS

SUBMITTED TO THE PROGRAM IN SYMBOLIC SYSTEMS

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

BACHELORS OF SCIENCE WITH HONORS

David Leo Wright Hall

May 2008

© Copyright by David Leo Wright Hall 2008  
All Rights Reserved

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a thesis for the degree of Bachelors of Science with Honors

---

(Daniel Jurafsky  
Linguistics and Computer Science) Principal Adviser

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a thesis for the degree of Bachelors of Science with Honors

---

(Chris Manning  
Computer Science and Linguistics)



# Abstract

I describe a series of computational techniques that can be used to analyze the influence and origins of scientific paradigms. These techniques are based on the automatic processing of papers in scientific papers and journals. First, I present a novel topic model that can automatically extract the vocabularies used by different paradigms and the degree to which certain authors and their works are affiliated with those paradigms. I then develop several different techniques to answer specific questions about the history of science: who is responsible for scientific revolutions; what becomes of members of the previous paradigms; whether two different paradigms merge; and what role interdisciplinarity plays in the formation of new ideas. I apply these tools to the field of Computational Linguistics in an attempt to understand the so-called “Statistical Revolution” in the early 1990’s that brought modern Artificial Intelligence techniques to the field. I present evidence to help determine what group of researchers was most responsible for initiating the paradigm shift, and examine the role that other groups of researchers played in the adoption of a new paradigm. The outcomes provide new insights about the history of the field. I then discuss the implications of the application of computer science methodology to the problems and methods of the social sciences.

# Acknowledgments

First, thanks to my parents, as always *sine quibus non*, for doing all the essential things that parents do (and being really good at it), but particularly for putting up and even encouraging my little projects, and also for that little thing called tuition. Then, to my thesis advisors Dan Jurafsky and Chris Manning, for patience, especially in getting me started so many different times, for the truly insightful questions and advice, for letting me get into my own little world of plate diagrams and parallelism, and then especially for pulling me out of it often enough so that I didn't lose the sight of the big picture too often. To Ivan Sag, for the great conversations—advising and otherwise—and for keeping me an honest Symbolic Systems major by reminding me to take more linguistics (the fun stuff). To Dan McFarland, for his incredible ability to ramp up the Minerva project and keep everyone excited and organized and for insights from outside the world of computational linguistics. To Daniel Ramage, Jenny Finkel, Sharon Goldwater and Anna Rafferty for the advice and good times in the Variational Reading group. Maybe we'll get it someday. To all the members of The Minerva Project and the NLP group, for all the interesting meetings, insightful questions and for listening to all the early versions of what's in here.

More operationally, thanks to Hal Daumé for putting up with my constant questions about his Hierarchical Bayesian Compiler, which was essential for early prototypes in this project. Drago Radev and Bryan Gibson, for so helpfully providing access to the data.

Finally, thanks to my friends who are thankfully too numerous to enumerate here, for putting up with my obscure questions, my strange status messages, and being there at all times.

# Contents

<b>Abstract</b>	<b>v</b>
<b>Acknowledgments</b>	<b>vi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Paradigms and Ideas . . . . .	2
1.2 Topics and Ideas . . . . .	3
1.3 Case Study: Computational Linguistics . . . . .	4
1.4 Roadmap . . . . .	8
<b>2 Preliminaries</b>	<b>9</b>
2.1 The vector space model . . . . .	9
2.1.1 Understanding distance and similarity . . . . .	10
2.1.2 Latent Semantic Analysis . . . . .	11
2.2 Mixture models and latent topics . . . . .	12
2.2.1 Probabilistic Latent Semantic Analysis . . . . .	12
2.2.2 The Simplex Model . . . . .	13
2.3 Generative models . . . . .	14

2.3.1	Latent Dirichlet Allocation . . . . .	14
2.3.2	Plate diagrams . . . . .	15
<b>3</b>	<b>Related Work</b>	<b>17</b>
3.1	Topic Dynamics . . . . .	18
3.2	PageRank . . . . .	20
3.3	Modeling Content and Citations . . . . .	21
<b>4</b>	<b>Ideas and The Structure of Normal Science</b>	<b>26</b>
4.1	Ideas as Topics . . . . .	26
4.2	The Family Tree of Science . . . . .	27
4.3	Standing on the Shoulders of Giants . . . . .	28
4.4	IdeaRank . . . . .	32
<b>5</b>	<b>Methodology</b>	<b>33</b>
5.1	The Corpus . . . . .	33
5.2	Computational Linguistics: The Hypotheses . . . . .	34
5.2.1	The Birth of an Idea . . . . .	36
5.2.2	The Fate of Ideas . . . . .	36
5.2.3	Paradigm Merger: Gazdar . . . . .	38
5.3	Model Estimation . . . . .	38
5.3.1	Collapsed Gibbs Sampling . . . . .	38
5.3.2	IdeaRank . . . . .	40
<b>6</b>	<b>Results and Discussion</b>	<b>41</b>
6.1	Topics . . . . .	41

6.2	Computational Linguistics: The Answers . . . . .	43
6.2.1	Paradigm Shift: The Rise of Probability . . . . .	43
6.2.2	The Decline of Unification . . . . .	49
6.2.3	Paradigm Merger? . . . . .	50
6.3	IdeaRank . . . . .	52
<b>7</b>	<b>Conclusion</b>	<b>54</b>
7.1	The Role of Interdisciplinarity . . . . .	54
7.2	Future Work . . . . .	55
	<b>Bibliography</b>	<b>57</b>



# Chapter 1

## Introduction

How do ideas shape future research? How do they become incorporated, changed, and expanded? In some sense, we have an intuition: “good” ideas get borrowed by other researchers and improved upon. These ideas, in turn, become borrowed and improved. This, goes the classical reasoning, is how science progresses.

However, Kuhn [1962] overturned this old view of scientific progress. In his monograph, Kuhn establishes that science truly progresses not by iterative tinkering but by revolutions in the tools and frameworks that scientists have at their disposal. These frameworks, called *paradigms* fundamentally shape the way most scientists operate. A paradigm not only shapes the models and ideas that are developed, but also the ways they collect data, what is considered data, and, perhaps most essentially, what kinds of questions can be asked at all.

This thesis establishes a method for determining the structure of a scientific revolution. By employing techniques from natural language processing, integrating techniques from computer science with the social sciences, this thesis seeks to measure the birth and death of the paradigms mentioned by Kuhn and test hypotheses from the history and philosophy of science. The framework developed here is applied to one particular revolution in the Computational Linguistics community known as the Statistical Revolution. In the span of just five years, the use of statistics and probability in Computational Linguistics went from fringe idea to dominant paradigm. Who was responsible for the paradigm shift, and what role did the meeting of two disciplines play? What became of the

researchers of the old paradigm? And can the two paradigms merge?

More generally though, this thesis hopes to show that the models and techniques developed in computational linguistics should be essential tools for the social scientists. The rich statistical information about language extracted from models like those presented here should prove useful in the design and construction of social and historical models. And, by the same token, I hope to demonstrate to computational linguists that the problems posed by social scientists provide interesting challenges that can help drive forward new research and applications.

## 1.1 Paradigms and Ideas

Sir Isaac Newton famously declared in a letter to Robert Hooke that “If I have seen further, it is by standing on the shoulders of Giants.” Newton’s pronouncement is typically regarded at once as a statement of extreme modesty and also a recommendation for how it is that science should progress. However, according to Kuhn [1962], what Newton saw—Newtonian dynamics—was different in kind from what the “Giants” before him saw. Newton’s most recent predecessors were engaged in *normal science*, the iterative advancement of science that remains inside a particular paradigm. Normal science results in the discovery of a series of *anomalies*: discoveries that are difficult or impossible to characterize satisfactorily in the present paradigm. At some point, some scientist (in this case Newton) looks at the sheer mass of anomalies and then fundamentally rethinks what they mean, and then develops a completely new framework that at a stroke resolves the anomalies before him.<sup>1</sup> This is the birth of a new paradigm, and this is the structure of scientific revolutions.

What is most important about paradigms in the Kuhnian sense is their *incommensurability*. Two paradigms fundamentally cannot be reconciled with each other. Different paradigms not only differ on how they interpret data from the world around them: they do not agree on what should be considered data. For instance, in Aristotelian dynamics, a pendulum was conceived of as “constrained fall,” and the description of its motion should be interpreted as the mass’s attempt to fall to Earth. However, in the Newtonian and Galilean analysis, a pendulum’s motion is (approximately) how I understand a pendulum today: the oscillation of an object with period

---

<sup>1</sup>Or her. In this thesis, I alternate across paragraphs between using the masculine and the feminine pronoun for a generic person.

proportional to the (square root of the) length, and not to the mass at all.

The depth of the incommensurability of paradigms goes even further. Members of one paradigm cannot logically argue for their paradigm over another, because any argument necessarily relies on the acceptance of the paradigm's goals and principles. Members of two differing paradigms see different worlds, and therefore one cannot argue that a new paradigm (Newtonian dynamics) better solves the problems of an earlier one (Aristotelian dynamics) better than the early one: they seek to solve different problems and have fundamentally different ways of talking about the problems they seek to solve. Similarly, the ideas generated by researchers engaged in "normal science," and the vocabularies used to describe those ideas, are fundamentally constrained by the paradigm they are operating under.

## 1.2 Topics and Ideas

What then is the measure of an idea? To a first approximation, a scientific idea is at once an approach to science and a way of describing that approach. That is, all ideas are characterized by a vocabulary. In fact, for Kuhn [1962], incommensurability arises precisely because the vocabularies used to describe the ideas in two paradigms are irreconcilably different. However, because different paradigms are necessarily connected to different vocabularies, I propose that to determine the vocabularies used to talk about paradigms may be sufficient to identify them.

How then can one separate one vocabulary from another? Most current approaches concern themselves with "latent topics," which are best thought of as particular vocabularies shared across papers [Hofmann, 1999]. Each paper's content is made up of a mixture of these vocabularies. For example, a document's content could be one-quarter about "sports" and three-quarter's about "news", while another could be one-quarter about "news," one-quarter about "weather," and one-half about "entertainment". It is important to note that the labels assigned to these topics are based on human judgments: strictly there is no "sports" topic. Instead, the name of the topic was assigned by me, while the topic itself is merely a distribution over possible vocabulary items, with a strong bias for using words like "baseball," "helmet," "yard," "championship," and "team," and a label would be ascribed later.

More specifically, this thesis will demonstrate that a paradigm can be represented by one or more latent topics that the models I develop extract. Similarly, an idea can be represented by a more narrow, less prominent topic that does not dominate a particular era. For instance, a paradigm from the Newtonian revolution might be represented by a topic best labelled “Empiricism”, while an idea may involve the specifics of Newtonian optics.

By determining the vocabularies (topics) used in a corpus of papers, and which papers employ which vocabularies, I hope to trace the origin of these vocabularies as well. And, if the vocabularies are specific enough, then it may be that new vocabularies emerge precisely where new ideas do. Therefore, to determine the history of ideas—and which papers are ultimately responsible for those ideas—one must study their vocabularies.

### 1.3 Case Study: Computational Linguistics

This thesis focuses on the development of the field of Computational Linguistics, from the 1970’s until the present. Computational Linguistics (also referred to as Natural Language Processing or NLP) can be described as the boundary between linguistics, artificial intelligence and psycholinguistics.

While this direction may seem like so much navel-gazing, there are several reasons why the choice of Computational Linguistics is ideal. First, there is the unprecedented availability of the entire corpus of scientific research in an entire field. Second, NLP is a reasonably large domain that contains a number of distinct subtasks, while on the other hand it is not too large: there are roughly 12,500 papers published by the Association of Computational Linguists (ACL). Third, there are several distinct subfields in NLP, including Automated Speech Recognition, Text-to-Speech, Information Retrieval, Machine Translation, Parsing, and Text Summarization. While distinct, these tasks borrow ideas from one another, and researchers often work in more than one of these subfields. For example, in recent years, members of the Machine Translation sub-community have been influenced by work in the Parsing sub-community. Fourth, there have been a few studies on paradigm shifts in Natural Language Processing, and combined with anecdotal evidence I can validate the methods on the covered phenomenon while making predictions in places with less study. And, finally, there is some navel-gazing involved too.

By determining the various vocabularies (topics) used in a field, and which vocabularies are used by which papers, I can then examine how papers borrow vocabularies from their predecessors, and whether or not these topics form paradigms.

## The Questions

What questions can we ask about computational linguistics? Researchers have already addressed some of these issues qualitatively; none have tried to give the history a quantitative analysis.

## Interdisciplinarity and the Origin of Ideas

Where do new ideas come from? In a Kuhnian analysis, new paradigms are undertaken by certain of the braver scientists: the Copernicuses, the Newtons, and the Einsteins. These researchers, having worked in their field for some time, identify a series of anomalies that make working within the existing paradigm increasingly difficult. In the 1980's, text-based computational linguistics was dominated by approaches that Gazdar [1996] labelled "LOGIC". These approaches followed the spirit of the original Chomskian revolution by focusing on representing language as part of a formal mathematical framework, notably certain variants of mathematical logic. Moreover, they were characterized by highly specific fine-tuning designed to capture the representation of an area of language completely. However, the very precision that the LOGIC approach emphasized led to increasing frustrations with its ability to capture general aspects of language. And so, beginning in the 1990's, and increasingly quickly, methods relating to statistics, which are characterized by high coverage but lower precision, took hold. In a Kuhnian framework, this is precisely the origin of a scientific revolution.

Historically, researchers in understanding spoken language have employed probabilistic methods: the quality of the source changes and humans are not consistent in the way they speak, in effect forcing stochasticity on any model of speech processing. However, perhaps surprisingly, speech- and text-based NLP are largely seen as distinct fields. Until the 1990's—and even since then—few researchers consider themselves both "speech" and "text" people [Gazdar, 1996]. However, in 1989, a series of workshops held by the ACL (a text-oriented association) included both speech and text papers. At about this time, probability suddenly exploded in the text-based community. Were

these somehow related?

If this cross-pollination with speech did indeed ignite the explosive growth of probabilistic methods in text-based NLP, then this challenges Kuhn: revolutions are not started by “brave” but experienced scientists within the field, but by new scientists from the outside looking to apply their paradigm to a new area. That is, at least sometimes, scientific revolutions depend on interdisciplinary research, and scientists looking to branch out into a new field.

### **The Fate of an Idea**

What becomes of old ideas? As Kuhn [1962] says through Max Planck, do they simply die?

“A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.”

Or, do they somehow become integrated with the new? In the Kuhnian world-view, new paradigms are incommensurable with the old: ideas die out precisely because they are old. The ideas of one paradigm cannot be used in another. Therefore, two paradigms cannot coexist. One must supplant another.

However, in the NLP community, where there has been around fifteen years since the paradigm shift, the researchers involved have not “died,” but have continued to publish. In particular, we focus on the parsing subcommunity within NLP. “Parsing” refers to the task of automatically extracting the constituent components of the sentence. For instance, given the sentence, “Pat ran down the street,” a parsing program would automatically determine that “ran” is the main verb, that “Pat” is the subject, that “down the street” is a prepositional phrase that modified “ran,” and several other more low-level facts about the sentence.

During the 1980’s and early 1990’s, the parsing community was dominated by an approach based on the “unification” of feature structures, which can be thought of as syntactic and lexical properties of words and combinations of words. For instance, Kay [1984] introduced Functional Unification Grammar, in which every word was associated with a kind of function, and phrases were built by “unifying” the constraints specified by the smaller parts. In a simplified version, the

verb “ran” might be associated with the logical properties of running, and also that “ran” should be accompanied by a subject. Similarly, “Pat” would be represented by a feature structure specifying that the person is Pat and that the word can be a subject or an object. Then, to form the sentence “Pat ran,” one would unify the structures by ensuring that the propositions specified by “ran” (that it takes a subject) and by “Pat” (that it can be a subject) are compatible. The unified structure then expresses the logical propositions encoded in the structures for “ran” and “Pat”, along with the extra proposition that Pat did the running. The paradigm represented by “Unification” was largely concerned with the representation and construction of these kinds of feature structures.

However, beginning in the mid-1990’s, the Unification paradigm, as such, seems to have almost entirely fallen from prominence, but the researchers themselves largely have not exited the field. Can one detect what happened to Unification? Did it transform into a new idea, or was it in fact supplanted by a new paradigm? And the researchers: did they embrace the new paradigm, or does Max Planck’s observation hold even in Computational Linguistics?

### **The Statistical Revolution: Paradigm Merger?**

Gazdar [1996] describes what he terms a “paradigm merger” beginning in the 1990’s of two wholly separate communities from the 1980’s. The first he labels “LOGIC” which is characterized by large, carefully hand-built systems involving such terms as “feature structures,” and “Unification,” together with hand-structured AI-based theorem provers. The “NGRAM” community was focused on gathering large amounts of data and using coarse-grained statistical approaches to process language mostly automatically. Both of these paradigms suffered from serious flaws: the LOGIC group, while achieving very high accuracy, involved great cost in specialists’ time and did not result in very high-coverage systems. The NGRAM approach required far more data, but the data did not need the intervention of specialists. The resulting systems were characterized by high-coverage but poor accuracy.

NLP practitioners, Gazdar reasons, realized that these two communities could benefit from each other. In particular, he points to signs that the Parsing community (historically a part of the LOGIC community) began to use the statistical approaches of the NGRAM community: researchers began to train statistical systems on specialist-annotated “treebanks” (e.g. the Penn treebank: Marcus

et al. [1994]).

Gazdar’s hypothesis is in direct opposition to Kuhn’s. If Gazdar is correct, then paradigms are not incommensurable, indicating that somehow researchers are able to bridge the gap in communication, and therefore mergers are indeed possible.

However, other anecdotal evidence suggests that the NGRAM community largely did not exist before the 1990’s, and when it arose, it quickly overtook the then-dominant LOGIC community. If there was no merger, then the vocabularies for these different paradigms should remain wholly distinct. Otherwise, the vocabularies must also somehow merge.

## 1.4 Roadmap

This thesis consists of six further chapters. Chapter Two introduces the background information necessary to understand the kinds of techniques and models that researchers have proposed to analyze questions like those proposed here. Chapter Three presents several of these models and the benefits and drawbacks of each. Chapter Four presents a novel model that directly models the progress of “normal science” for determining how scientists borrow ideas from their predecessors. Chapter Five discusses the methodology used to analyze the questions posed in the Introduction, while Chapter Six presents and discusses the results. Chapter Seven, the conclusion, discusses the implications for the results presented in Chapter Six as well as discuss directions for future research.

# Chapter 2

## Preliminaries

Fundamentally, this thesis is concerned with the extraction of different vocabularies that should represent different paradigms and ideas as defined by Kuhn [1962]. This techniques in this thesis employ variants of document clustering methods to extract vocabularies: they all rely on determining the similarity of documents automatically. In order to determine that similarity, they all assume mutually circular notions of document and word similarity: similar documents have similar words, and similar words are found in similar documents. All of these techniques further rely on certain mathematical representations of documents, which are described in this chapter. The treatment given here is designed so that the first two or three sections are designed to give the intuitions (and some details) behind the representations, while the later sections are more mathematically detailed, though even those try to focus on intuitions rather than rigor.

### 2.1 The vector space model

The models I discuss in this thesis have a history rooted in the *Vector Space Model*, in which one represents documents as vectors of word counts [Manning et al., 2008]. For example, consider three very short documents:

1. *Words are the physicians of a diseased mind.* – Aeschylus

2. *Words, words, words.* – Shakespeare

3. *Words are all we have.* – Beckett

Moreover, assume that all the vocabulary of my language is limited to the 11 distinct word types found in these three documents. Then, in the vector space model, one could represent these documents as vectors in this 11 dimensional space:

Key	words,	are,	the,	physicians,	of,	a,	diseased,	mind,	all,	I,	have
Aeschylus	1,	1,	1,	1,	1,	1,	1,	1,	0,	0,	0
Shakespeare	3,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0
Beckett	1,	1,	0,	0,	0,	0,	0,	0,	1,	1,	1

A few observations are in order. First, each entry corresponds to a dimension of the vector space. The more one sees the word “physician” in a document the further along the physician-dimension the document is. Second, the vector space model exhibits what is alternatively called the *bag of words* or *exchangeability* assumption: the order of the words does not affect the meaning of the vector. That is, one can rearrange Aeschylus to say “Physicians are a mind diseased of the words.” and it would still have the same representation in the vector space. Ultimately, this means that these models are best for high level content. The nuance of syntax is completely lost, but the models are far more tractable computationally.

### 2.1.1 Understanding distance and similarity

By appealing to a mathematical representation of documents, we can also consider how similar (or different) two documents are. Intuitively, the more different two document vectors are, the less similar the actual documents are. That is, the more words two documents share (i.e. the more they are represented in the same dimensions), the more similar they are. Therefore, one might represent the similarity of two documents to be the *dot product* of their document vectors:

$$d_1 \cdot d_2 = \sum_{i=1}^V C_{d_1, w_i} \cdot C_{d_2, w_i} \quad (2.1)$$

where I define  $C_{d,w}$  to be the number of times word  $w$  appears in document  $d$  and  $V$  to be the size of the vocabulary. Using this formula, the similarity of these documents are:

$$d_A \cdot d_S = 3$$

$$d_A \cdot d_B = 2$$

$$d_S \cdot d_B = 3$$

An artifact of the dot product is that one can make the value go up just by repeating the document over and over again; if Hamlet had repeated “words” thirty times instead of simply three, then the scores would have increased ten-fold. Of course, this is not particularly useful, so instead the *cosine* is often used, which measures the size of the angle between the two vectors. In practice, it is exactly the same as the dot product, except that one truncates the vectors so that they have length one. That is, we are only interested in the direction of the vector, and not its magnitude. Cosine distance is therefore defined as:

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|} \quad (2.2)$$

, where  $\|d\| = \sqrt{\sum_i C_{d,w_i}^2}$ . By this measure, we have similarities:

$$\cos(d_A, d_S) \approx .35$$

$$\cos(d_A, d_B) \approx .32$$

$$\cos(d_S, d_B) \approx .45$$

### 2.1.2 Latent Semantic Analysis

However, both of these metrics have a substantial problem. Ideally, I would like the quotation “Words are those doctors for a sick brain” to be almost identical to “Words are the physicians of a diseased mind”. However, the current representation does not support the notion of word similarity. To solve this, several authors have proposed the *Distributional Hypothesis*: words that have similar

meanings appear in similar contexts [McDonald and Ramsar, 2001]. In particular, Deerwester et al. [1990] propose Latent Semantic Analysis (LSA), which finds a lower dimensional approximation to the vector space by merging dimensions that appear similar. For instance, “doctor” and “physician” may not be found near each other often, but they appear with “medicine” and similar words quite often, and thus they might be conflated in the lower dimension. That is, I examine the distributions of words; if they are similar, then in a lower dimension I might just approximate them as the same word. This low dimensional approximation then represents each document as a vector along these new merged dimensions, and therefore I can use the same measures as from the previous section.

## 2.2 Mixture models and latent topics

One can also think about a document as a mixture of several different vocabularies: a biology paper may contain words relating to statistical analysis, some kind of experimental procedures, background about DNA, and about photosynthesis—to arbitrarily name several possible vocabularies. In particular, I can conceive of the authors of a document having proportions for how much certain vocabularies should be represented. For instance, the author of that biology paper could have intended to focus strongly on photosynthesis (say, about 70% of the paper) and the other 30% would be divided evenly among the other 3 vocabularies.

If we think about documents in this way, then the task is to discover not only what proportions each document has for the vocabularies, but also what the vocabularies are. These vocabularies are represented by what are called *latent topics* or *aspects*. These names hint at their role: the topics themselves are not readily apparent in the document, but instead represented only by the words from their vocabularies.

### 2.2.1 Probabilistic Latent Semantic Analysis

To find both topics and proportions, Hofmann [1999] proposed Probabilistic Latent Semantic Analysis (pLSA). pLSA represents word  $w_{ji}$  in document  $d$  by considering the probability that the word belongs to the  $k$ 'th topic  $z_k$ , multiplying that by the proportion with which document  $d_j$  uses topic

$z_k$ , and then summing over all possible topic assignments:

$$P(w_{ji}|d_j) = \sum_k^K P(w_{ji}|z_k)P(z_k|d_j) \quad (2.3)$$

These probabilities can be estimated by iteration using the Expectation-Maximization algorithm [Hofmann, 1999].

## 2.2.2 The Simplex Model

I can view pLSA much like LSA, except that now I am finding a  $K$ -dimensional approximation to the words in each document, where  $K$  is the number of topics. More specifically, each document has a probability distribution  $p(z|d_i)$  which is a multinomial distribution, and thus representable as a vector of probabilities for each of the topics in the document. These vector components must sum to 1, so strictly the approximation requires all the documents to be on the  $(K-1)$ -simplex. Distance on the simplex can be measured in the same as with words and latent dimensions, but other metrics are more common. These formulae measure the *divergence* between two probability distributions, that is, the difference in the probabilities they assign to their possible values. The most commonly used is the *Kullback-Leibler (KL) divergence* which, for two probability distributions  $P$  and  $Q$  is given by: [Manning et al., 2008]

$$D_{\text{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad (2.4)$$

The KL-divergence is always non-negative, but if  $Q(x) = 0$  for any  $x$ , then the  $D_{\text{KL}}(P||Q) = \infty$ . Moreover, the KL divergence is not symmetric:  $D_{\text{KL}}(P||Q) \neq D_{\text{KL}}(Q||P)$  for most distributions. Because of these features, it is often useful to consider the distance between their average distribution: the *Jensen-Shannon Divergence* or the *Information Radius*:

$$\begin{aligned} D_{\text{JS}}(P||Q) &= \frac{1}{2}(D_{\text{KL}}(P||R) + D_{\text{KL}}(Q||R)) \\ R &= \frac{1}{2}(P + Q) \end{aligned} \quad (2.5)$$

The Jensen-Shannon divergence is always finite and symmetric, making it a better choice for certain kinds of distributions and applications.

## 2.3 Generative models

Probabilistic Latent Semantic Analysis, however, does not provide a means of expanding to more documents. Because it represents a distribution  $p(z|d)$ , pLSA is limited to the number of documents it was initially built on. Instead, one would prefer to consider models that can be easily extended to new documents. Such “generative” models are described by *generative stories* that describe the origin of the data. Generative stories are intentionally not accurate reflections of how the data is actually created. Instead, they are simplifications (or even fabrications) to make the statistical models based on these stories feasible to train. However, they can often reveal interesting properties about the data, even when they are not accurate.

The models I consider in this thesis focus on the process by which authors generate documents. Typically, an author will choose how many words long her document will be, and then for each of those words, she will choose which topic she wants the word to be about, and then she will choose a word from that topic.

### 2.3.1 Latent Dirichlet Allocation

In fact, most of the models considered in this thesis are descendants of a particular model: Latent Dirichlet Allocation (LDA; Blei et al. [2003]). LDA is a fully generative version of probabilistic Latent Semantic Analysis that models (a version of) the process by which documents are created. Instead of representing  $p(z|d)$ , we introduce a new parameter  $\vec{\theta}_d$  for each document, which represents the proportion of the words in each document belong to the various topics. Strictly,  $\theta_d$  is the parameter for a multinomial distribution:  $\theta_{d,z}$  is the probability that an arbitrary word in document  $d$  is topic  $z$ . Similarly, we write  $p(w|\beta_z, z)$  instead of  $p(w|z)$ , and treat  $\beta_z$  as the multinomial distribution over words (i.e. the vocabulary) for topic  $z$ .

Putting this together, the generative story deliberately reads much like an algorithm for writing

$$\begin{aligned}
 D &= \text{the number of documents in the corpus} \\
 K &= \text{the number of topics} \\
 N_i &= \text{the number of words in document } i \\
 \theta_i &\sim \text{Dir}(\alpha) \\
 \beta_k &\sim \text{Dir}(\eta) \\
 z_{di} &\sim \text{Mult}(\theta_i) \\
 w_{di} &\sim \text{Mult}(\beta_{z_{di}})
 \end{aligned}$$

Figure 2.1: Specification for Latent Dirichlet Allocation

papers:

- For each document  $d$ :
  - Choose  $\theta_d$ , the topic proportions for document  $d$ , from the parameters  $\alpha$ .
  - For each word  $w_{di}$  in the document:
    - \* Choose  $z_{di}$ , the topic for the  $i$ 'th word, from  $\theta_d$ .
    - \* Choose  $w_{di}$  from the vocabulary distribution  $\beta_{z_{di}}$ .

Just as with probabilistic Latent Semantic Analysis, one can think of Latent Dirichlet Allocation as a probabilistic version of dimensionality reduction. Specifically,  $\theta_d$  can be understood as an approximation of document  $d$ 's content on the  $(K-1)$ -simplex. (Recall that  $K$  is the number of topics  $z$ ). For completeness, the formal specification for Latent Dirichlet Allocation is in Figure 2.1.

### 2.3.2 Plate diagrams

A common way of expressing generative stories is through the use of *plate diagrams*, which represent the repeated processes used.<sup>1</sup> For example, the plate model for LDA is in Figure 2.2. Each circle represents a variable or a parameter, and the arrows denote influence. Shaded variables represent

---

<sup>1</sup>It should be noted that plate diagrams can be used for non-generative models. However, they become slightly more difficult to understand outside of the generative idiom.

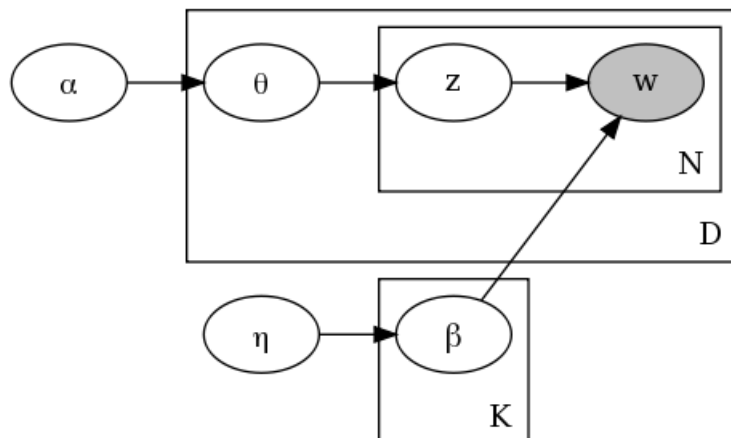


Figure 2.2: Plate Model for LDA

those variables that are observed. For instance, the words (the  $w$ 's) are observed, and the arrow from the  $z$  to the  $w$  means that the choice of topic for a specific word directly influences the choice of word. Each plate (box) should be thought of as specifying “each” or “for each,” and the number in the lower right hand corner represents the number of copies of that plate. Thus, the outermost plate with the  $D$  in the lower right hand corner represents documents corresponds to the outermost loop in the generative story. The inner plate represents the second for-each loop, and the other smaller plate indicates that there are  $K$  copies of the vocabulary distributions, one for each topics. (This last plate should be thought of as carrying the “each” meaning.)

# Chapter 3

## Related Work

In the previous chapter I introduce some basic models for automatically learning the vocabularies that different scientific paradigms and ideas within those paradigms employ by clustering the vocabularies of latent topics automatically. However, these basic models do not explicitly represent time or the impact that previous documents have on future documents. Both of these features seem relevant to the Kuhn [1962] thesis: we need both a notion of prominence over time to understand the rise and fall of paradigms and a representation of the “normal science” that proceeds by iterative refinement of previous ideas.

Therefore, we review new models have explicitly modeled time by encoding the popularity that certain topics had at different points in history. Others have modeled the connection between documents that cite the same papers. Most of these models can be thought of as extensions to Latent Dirichlet Allocation (LDA; [Blei et al., 2003]), which was introduced in the previous chapter. This chapter reviews this work on topic dynamics and versions of Latent Dirichlet Allocation that include citation information. I also review the specifics of the PageRank[Page et al., 1998] algorithm. These models provide the necessary background to understand the novel models presented in Chapter Four by placing it in the tradition of topic modeling that has evolved over the past several years.

### 3.1 Topic Dynamics

A small but increasing body of work has focused on *topic dynamics*. That is, how do topics, and the words used to describe them, change over time? While it is possible to analyze at least the question of topic prominence *post hoc*, researchers have tried to incorporate the knowledge directly into their models. Blei and Lafferty [2006] start from the intuition that the prominence of topics should shift only slightly from time period to time period, taking “steps” between periods. Their Dynamic Topic Model (DTM) represents documents as points on the topic-distribution simplex about a mean which can be thought of as the “zeitgeist” for the given epoch. Moreover, each epoch’s centroid is sampled from the previous year’s centroid, enabling “drift” of topics from epoch to epoch. Similarly, the vocabulary distributions for each of the topics are chosen as a centroid. Their generative story is something like the following:

- For each epoch  $t$ :
  - Sample  $\mu_t$ , the mean for this epoch, from a normal distribution centered around the previous epoch.
  - For each topic  $z$ :
    - \* Choose  $\vec{v}_{t,z}$ , the vocabulary distribution for topic  $z$  in time  $t$ , from a normal distribution centered around  $\vec{v}_{t-1,z}$ .
- For each document  $d$  in time epoch  $t$ :
  - Sample  $\theta_d$ , the topic proportions for this document, from  $\mu_t$ .
  - For each word  $w_{di}$ :
    - \* Sample a topic  $z_{di}$  from  $\theta_d$
    - \* Sample a word  $w_{di}$  from  $\vec{v}_{t,z_{di}}$

As one can see, the generative story is not terribly different from Latent Dirichlet Allocation’s. At each epoch, the distribution that topics and words are sampled from changes. However, the Dynamic Topic Model employs a normal distribution and a normalization function to constrain the points to be on the simplex. The model is specified fully in Figure 3.1. This model requires

$$\begin{aligned}
T &= \# \text{ epochs} \\
D_t &= \# \text{ documents in the corpus for the } t\text{'th epoch} \\
K &= \# \text{ topics in the corpus} \\
N_{t,i} &= \# \text{ words in document } i \text{ for epoch } t \\
\alpha_t &\sim \text{Norm}(\alpha_{t-1}, \sigma_1 I) \\
\beta_t &\sim \text{Norm}(\beta_{t-1}, \sigma_2 I) \\
\theta_{t,i} &\sim \text{Norm}(\alpha_t, \sigma_3 I), i \in [1, D_t] \\
z_{t,i,n} &\sim \text{Mult}(\alpha(\theta_{t,i})), i \in [1, D_t], n \in [1, N_{t,i}] \\
w_{t,i,n} &\sim \text{Mult}(\alpha(\beta_{t,z_{di}}))
\end{aligned}$$

Figure 3.1: Dynamic Topic Model

a substantial change to the inference procedures from LDA as well as renormalizing topic and vocabulary distributions by a transform function  $\alpha$ .

Another model is Topics over Time[Wang and McCallum, 2006]. Instead of dividing time into epochs, Topics over Time assigns a time stamp  $t$  in the range  $[0,1]$  to each document and repeatedly samples the timestamp according to a topic-dependent Beta distribution  $\psi_z$ , which gives a smooth distribution. Representation of time as a continuous variable has several useful properties. First, it is a rather minimal addition to the model, making it easy to implement. Moreover, it avoids the problem of Markovization: there is no need to divide time into a specific number of epochs. This negates the need to determine the proper dividing line for documents and the need to decide whether or not to cluster natural epochs. Thus, Topics over Time's generative is rather different:

- For each document  $d$ :
  - Choose  $\theta_d$  from the parameters  $\alpha$ .
  - For each word  $w_{di}$ :
    - \* Choose  $z_{di}$ , the topic for the  $i$ 'th word, from  $\theta_d$ .
    - \* Choose  $w_{di}$  from the vocabulary distribution  $\beta_{z_{di}}$ .
    - \* Choose  $t_{di}$ , the time at which the document is written, from  $\psi_{z_{di}}$

$$\begin{aligned}
D &= \text{the number of documents in the corpus} \\
K &= \text{the number of topics in the corpus} \\
N_d &= \text{the number of words in document } d \\
\theta_i &\sim \text{Dir}(\alpha) \\
\beta_k &\sim \text{Dir}(\eta) \\
z_{di} &\sim \text{Mult}(\theta_i) \\
w_{di} &\sim \text{Mult}(\beta_{z_{di}}) \\
t_{di} &\sim \text{Beta}(\psi_{z_{di}})
\end{aligned}$$

Figure 3.2: Topics over Time

Note that, due to modelling constraints, the time stamp for each document is chosen repeatedly for every word. (Topics over Time also constrains the time stamp to be the same.) Otherwise, the generative story is precisely the same as Latent Dirichlet Allocation’s. Topics over Time is specified in Figure 3.2, and the plate model is in Figure 3.3.

## 3.2 PageRank

In a different direction, Page et al. [1998] propose PageRank, which models the behavior of a random surfer on the Internet. He starts randomly on some web page, and then arbitrarily chooses one of the links on the page to follow to a new page. However, sometimes (with constant probability) the surfer decides to not follow one of the links presented to him, but instead he randomly restarts on a new page. The process repeats until a steady state emerges, where the surfer has a fixed probability of being at a given page. The equation to describe this distribution is therefore:

$$PR'(d) = c \sum_{d':d \in C_{d'}} \frac{PR(d')}{|C_{d'}|} + \frac{1-c}{D} \quad (3.1)$$

where  $d$  and  $d'$  are documents,  $C_{d'}$  is the set of documents that  $d'$  cites,  $D$  is the total number of documents, and  $c$  is a “dampening factor” that represents the probability that the random surfer follows a link instead of randomly restarting. PageRank can be calculated by repeatedly evaluating

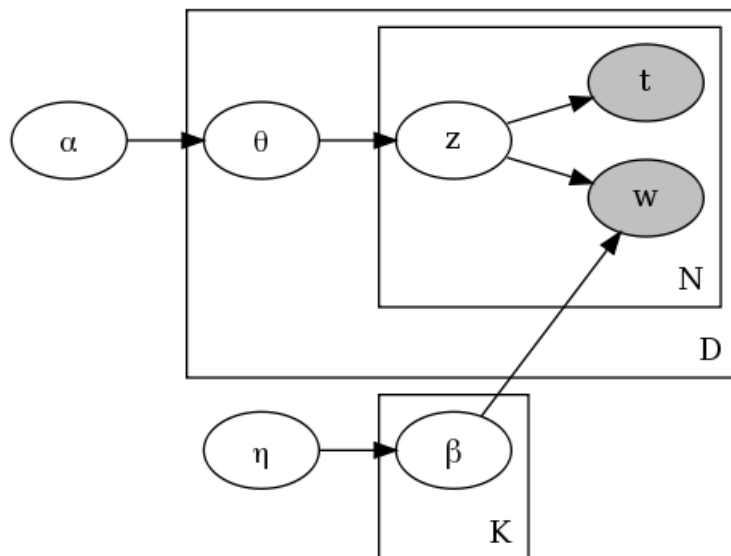


Figure 3.3: Topics over Time

this formula for each document  $d$  until convergence.

### 3.3 Modeling Content and Citations

Some recent work has considered the problem of integrating both topics and references. Cohn and Hofmann [2001] introduce pHITS, which can be thought of as probabilistic Latent Semantic Analysis that also has information about which documents cite which documents. In particular, one simply extends the original pLSA definition from Equation 2.3:

$$\begin{aligned}
 P(w_{ji}|d_j) &= \sum_k P(w_{ji}|z_k)P(z_k|d_j) \\
 P(c_{j\ell}|d_j) &= \sum_k P(c_{j,\ell}|z_k)P(z_k|d_j)
 \end{aligned}
 \tag{3.2}$$

, where  $c_{j,\ell}$  is the  $\ell$ 'th citation in the  $j$ 'th document, just like  $w_{j,i}$  is the  $i$ 'th word in the  $j$ 'th document. Thus, words and citations are treated in more or less the same way in pHITS. That is, the cited documents that the citations refer to have no defined relation to the original document. A

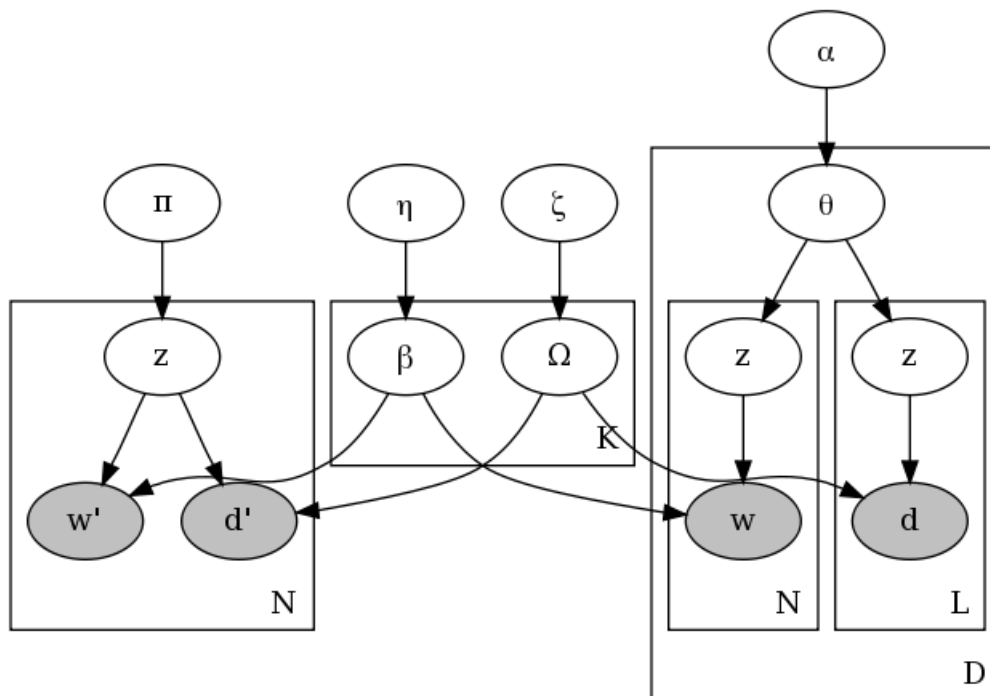


Figure 3.4: Link-pLSA-LDA

similar model based on Latent Dirichlet Allocation and not pLSA was proposed by Erosheva et al. [2004] that again treats words and citations in analogous fashions.

More recent work includes information about the cited documents. Nallapati and Cohen [2008] introduce Link-pLSA-LDA, which, as the name suggests, creates a link between Latent Dirichlet Allocation and probabilist Latent Semantic Analysis. In particular, it creates two copies of each document: one that cites, and one that is cited. The cited documents are treated as though they are in pLSA, and the citing documents are fully generative as in LDA. The plate model is in Figure 3.4. The pLSA part is on the left and the LDA component is on the right. The parameter  $\Omega$  constrains the distribution of both citing document and cited document to have similar topic distributions. Note that, because part of the model is based on pLSA, the model is not fully generative, and so does not generalize to new data with relearning the model.

Dietz et al. [2007] provide a fully generative analog of Link-pLSA-LDA called the Citation Influence Model (CIM; Figure 3.5). As in Link-pLSA-LDA, each document has both a cited and

citing version, though CIM provides generative semantics to both plates. The plate model in Figure 3.5, while initially quite complicated, can be understood as a two-phase generative process. First one generates the cited documents as in LDA. Then, one generates the citing documents by alternating between being original and borrowing from the cited documents (and in which proportions,  $\gamma_d$ ). To elaborate further:

- For each cited document  $d$ :
  - Choose  $\theta_d$ , the topic proportions for document  $d$ , from the parameters  $\alpha$ .
  - For each word  $w_{di}$  in the document:
    - \* Choose  $z_{di}$ , the topic for the  $i$ 'th word, from  $\theta_d$ .
    - \* Choose  $w_{di}$  from the vocabulary distribution  $\beta_{z_{di}}$ .
- For each citing document  $d$ :
  - Choose  $\gamma_d$ , the proportions with which document  $d$  borrows from the cited documents.
  - Choose  $\lambda_d$ , the original topic proportions  $d$ .
  - Choose  $\psi_d$ , the proportion of words that are novel.
  - For each word  $w_{di}$ :
    - \* Choose  $s_{di}$ , whether or not this word is original, from  $\psi_d$ .
    - \* If  $s_{di} = 0$ :
      - Choose  $r_{di}$ , the reference for the  $i$ 'th word, from  $\gamma_d$ .
      - Choose  $z_{di}$ , the topic for the  $i$ 'th word, from  $\theta_r$ .
    - \* Otherwise:
      - Choose  $z_{di}$ , the topic for the  $i$ 'th word, from  $\lambda_d$ .
    - \* Choose  $w_{di}$  from the vocabulary distribution  $\beta_{z_{di}}$ .

One might be concerned about the duplication of the documents, especially in the Citation Influence Model. Because the algorithms used are not exact, the learned topic proportions for the dual versions of cited and citing documents could, in theory, be substantially different. However, Dietz et al. [2007] found that in fact both versions were quite similar. Measuring the Jensen-Shannon

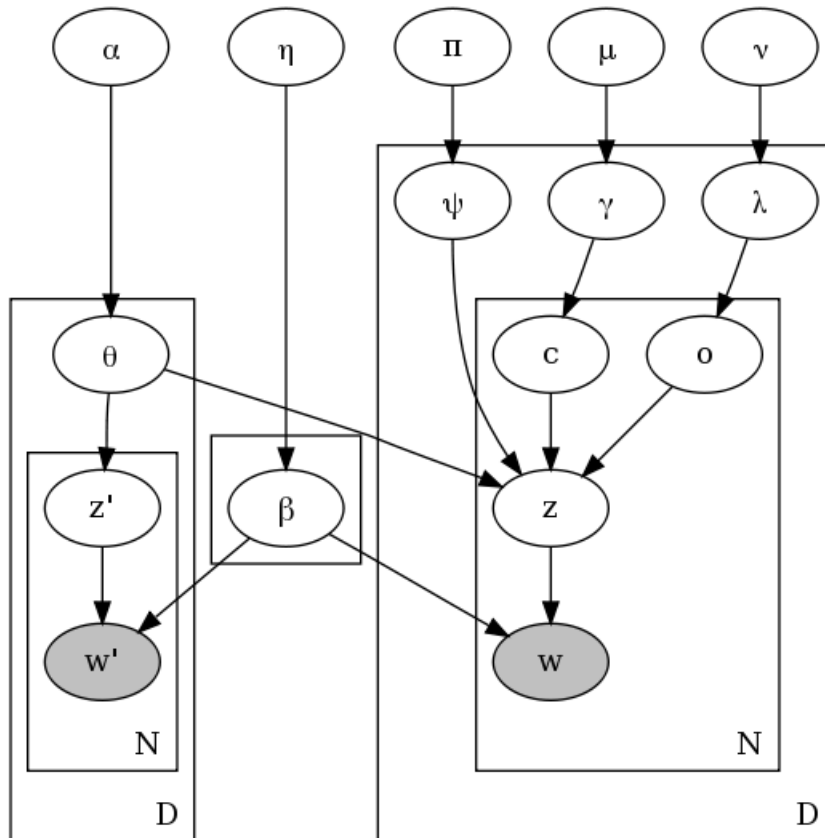


Figure 3.5: Citation Influence Model (CIM)

Divergence (see Equation 2.5) between the topic proportions for both cited and citing documents, they found an average of 0.07 between cited and citing documents, compared with 0.69 for arbitrary documents.

# Chapter 4

## Ideas and The Structure of Normal Science

In order to truly understand the role that paradigms shifts or revolutions play in science, one must understand how “normal science” is done. This chapter considers the usual progress of normal science: how to represent the ideas that scientists generate, how to determine which ideas scientists are improving upon, and which papers are the most influential for the functioning of ideas within a given paradigm. After considering these factors, I present a novel model to represent the history of a field, and present a modified version of PageRank to capture which papers are most influential.

### 4.1 Ideas as Topics

In the first chapter, I discussed Kuhn [1962]’s hypothesis about the incommensurability of different paradigms: a newer paradigm cannot be said to be better than an older paradigm because they attempt to solve fundamentally different questions. Based upon that hypothesis, I advanced the hypothesis that vocabularies, the words scientists use to discuss the ideas within a paradigm, are sufficient to characterize a paradigm.

But what of the ideas that remain within a paradigm? Thus far, I have (somewhat deliberately) left the word “idea” unspecified. It is, of course, inherently an abstract notion. One can say that

ideas are the things that scientists create and borrow from other scientists. However, it is not clear that a more rigorous definition would help. Instead, based on the Kuhn [1962] analysis from Chapter 1, it is sufficient to discover when a scientist has borrowed from another scientist. And for that, one needs a way to characterize the symptoms of idea transmission.

described precisely by the vocabulary used to describe it. That is, for the current approximation an idea is precisely the same as a latent topic. Thus, when a scientist adopts the vocabulary (a topic) from a paper she cites, this is equivalent to her borrowing that idea from that paper.

## 4.2 The Family Tree of Science

Conventional wisdom is that science proceeds in iterations. Each paper builds upon earlier papers, improving, fine-tuning, and merging ideas. These papers are in turn cited, and their ideas are borrowed, improved, and merged. Credit, in the Merton [1988] sense, builds in this way: better papers have better ideas, and so citation serves to indicate whose ideas are indebted to whose.

By this view, science can be modeled as a directed acyclic graph.<sup>1</sup> In particular, each paper is a node, and each paper has an edge to each of the papers it cites. Examining this graph on a macro scale, one can easily see that it is—in some sense—a family tree of science. This paper inherits certain traits from this paper it cites, and this paper inherits traits from this other paper, which in turns gets many of its ideas from another paper.

This metaphor can be extended further: we are interested in the history of the ideas, not just the papers that contain them. By analogy, we are more concerned with the genotypes of papers, rather than the phenotypes, the latent ideas rather than the apparent words.

---

<sup>1</sup>As a technical detail, I should note that in fact science is not quite acyclic. Occasionally papers published near the same time will cite each other. However, the principle remains largely true. The mathematics becomes more complicated, but we will see that in practice this metaphor works sufficiently well.

### 4.3 Standing on the Shoulders of Giants

To capture the history of these latent ideas, I propose new a class of models based on this directed (almost) acyclic graph: each paper is generated in part by the content of its “ancestors,” the papers that it cites. This class of models, which I call “Shoulders of Giants Models” after Isaac Newton’s famous quotation, which directly model the influence that scientific papers have on their intellectual descendants. These models differ from Citation Influence Model [Dietz et al., 2007] because they do not create both citing and cited versions of documents. Because science citations graphs are (almost) acyclic, I can simply specify that papers borrow directly from their ancestors, and are in turn borrowed from by their descendants. Thus, each paper is placed in the “family tree” at its place in history, and its topics are drawn based on its ancestors.

This thesis fleshes out the first of this new class of models, referred to simply as the Shoulders of Giants Model, is illustrated in Figure 4.1. Note that I abuse the notation: the dashed line from  $\theta$  to  $z$  should indicate that these are the topic proportions from the cited documents, analogous to the line from  $\theta$  to  $z$  across plate boundaries in Figure 3.5.

The generative story for the Shoulders of Giants Model reads much like a modification of the “citing” document portion of the Citation Influence Model.

- For each document  $d$ :
  - Choose  $\gamma_d$ , the proportions with which document  $d$  borrows from the cited documents.
  - Choose  $\lambda_d$ , the original topic proportions  $d$ .
  - Choose  $\psi_d$ , the proportion of words that are novel.
  - For each word  $w_{di}$ :
    - \* Choose  $s_{di}$ , whether or not this word is original, from  $\psi_d$ .
    - \* If  $s_{di} = 0$ :
      - Choose  $r_{di}$ , the reference for the  $i$ 'th word, from  $\gamma_d$ .
      - Choose  $z_{di}$ , the topic for the  $i$ 'th word, from  $\theta_r$ .
    - \* Otherwise:
      - Choose  $z_{di}$ , the topic for the  $i$ 'th word, from  $\lambda_d$ .

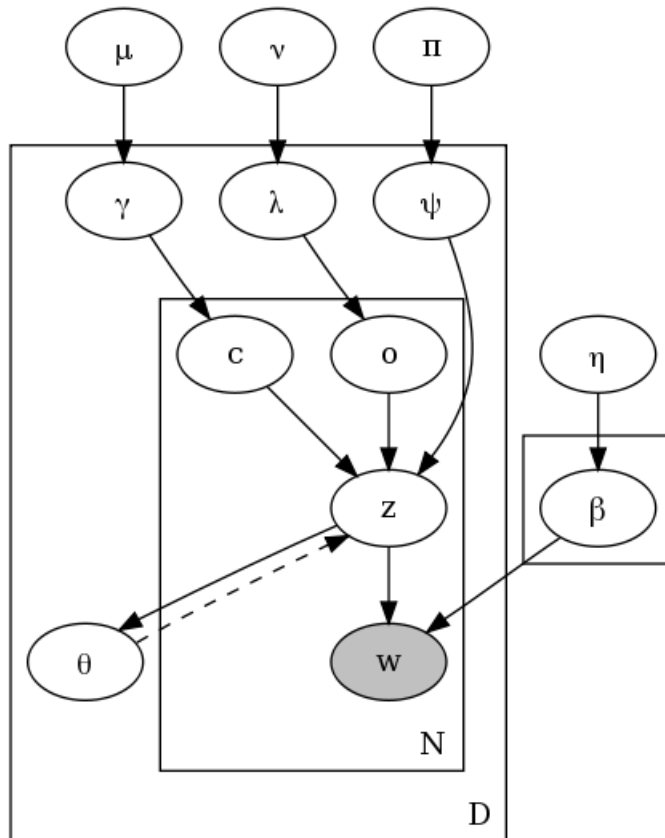


Figure 4.1: Shoulders of Giants Model (SGM)

$$\begin{aligned}
D &= \text{the number of documents in the corpus} \\
K &= \text{the number of topics in the corpus} \\
N_d &= \text{the number of words in document } d \\
\beta_k &\sim \text{Dir}(\eta) \\
\gamma_d &\sim \text{Dir}(\mu) \\
\lambda_d &\sim \text{Beta}(\vec{\nu}) \\
\psi_d &\sim \text{Dir}(\alpha) \\
o_{di} &\sim \text{Bin}(\lambda_d) \\
c_{di} &\sim \text{Bin}(\gamma_d) \text{ if } o_{di} = 0 \\
z_{di} &\sim \begin{cases} \text{Mult}(\psi_d) & \text{if } o_{di} = 1 \\ \text{Mult}(\theta_{c_{di}}) & \text{if } o_{di} = 0 \end{cases} \\
w_{di} &\sim \text{Mult}(\beta_{z_{di}}) \\
\theta_d &= \sum_i^{N_d} z_{di}
\end{aligned}$$

Figure 4.2: Shoulders of Giants Model

- \* Choose  $w_{di}$  from the vocabulary distribution  $\beta_{z_{di}}$ .
- Set  $\theta_d = \alpha_\theta + \sum_i z_{di}$ , the smoothed empirical Bayes estimate of the observed topics.

The formal specification is in Figure 4.2. A careful description of all the parameters is in order:

- $\eta$  : The parameter for the symmetric Dirichlet prior on the topic vocabularies.
- $\mu$  : The parameter for the symmetric Dirichlet prior on the citation proportionality multinomials.
- $\vec{\nu}$  : The parameter for the Beta prior on the originality constraints. Not symmetric to represent the prior belief about papers mostly building on prior ideas.
- $\alpha$  : The parameter for the symmetric Dirichlet prior for a document's original topic proportions, a multinomial over topics.
- $\beta_k$  : The vocabulary for the  $k$ 'th topic, a multinomial over words.

- $\gamma_d$ : The citation proportions for document  $d$ , a multinomial over cited documents.
- $\lambda_d$ : The originality probability for document  $d$ .
- $\psi_d$ : A document’s original topic proportions: a multinomial over topics.
- $o_{di}$ : Whether or not the  $i$ ’th word in the  $d$ ’th document is generated from the original proportions.
- $c_{di}$ : Which cited document’s topic proportions  $\theta$  the  $i$ ’th word is generated from when  $o_{di}$  is 0.
- $z_{di}$ : The topic for the  $i$ ’th word in document  $d$ .
- $w_{di}$ : The  $i$ ’th word in the  $d$ ’th document.

## Comparison of Two Models

The crucial difference between the Citation Influence Model and the Shoulder of Giants Model is that SGM requires only one copy of each document. As mentioned earlier, CIM documents are represented twice: once as the document that is cited and once as the document that does the citing. The two copies of the two documents are surprisingly similar, despite having only one weak link (through the vocabulary).

However, as the prior on the originality parameter grows to strongly favor “unoriginality,” the models differ significantly. In particular, the Shoulders of Giants Model will require that only those vocabularies which are borrowed by citing papers become extracted as topics at all, while the Citation Influence Model, or—as the authors call this degenerate case—the Copycat Model, will still generate largely the same topics. That is, with the right parameters, SGM can capture only those ideas that get picked up on by subsequent authors, which is, after all, the ideas that matter most.

## 4.4 IdeaRank

One of the core problems with the naïve versions of PageRank is that each link is treated equally. That is, a random reader of research papers is equally likely to transition to any of the cited papers. However, if I instead posit that a random reader is more likely to read papers that the original paper drew a great proportion of its ideas from, I can then use those proportions as an estimate of “idea-aware” PageRank transition probabilities. Using the learned citation proportions from the Shoulders of Giants Model  $\gamma_d$ , I can adapt 3.1:

$$IR'(d) = c \sum_{d':d \in C_{d'}} \gamma_{d'd} \cdot IR(d') + \frac{1-c}{D}$$

In this equation, I replace the uniform transition probability denoted by  $1/|C_{d'}|$  with the topic-borrowing probability  $\gamma_{d'd}$ .

# Chapter 5

## Methodology

This chapter seeks to answer the questions outlined in the introduction: who is most responsible for the “Statistical Revolution” in Computational Linguistics, what became of the researchers involved in the old paradigm, and did the two paradigms in fact “merge”? According to the paradigm-vocabulary hypothesis, I must first determine the different vocabularies used by the different paradigms. To do so, the models discussed in the previous chapter are applied to the research papers published by the Association of Computational Linguists.

### 5.1 The Corpus

The ACL Anthology is a publicly available collection of almost all of the publications in the journal *Computational Linguistics* and the conferences and workshops sponsored by the ACL. Currently, it comprises some 12,500 documents from conferences beginning as early as 1965 through 2006. Papers are organized by year and conference or journal. There are twelve different categories, and the number of papers and the ranges of their activity are listed in Table 5.1. Except for the Journal, all are conferences. There is considerable variety in the number of conferences, but a few should be mentioned in particular. The Journal generally is regarded with the highest esteem, and the main ACL conference is generally regarded as one of the most important. The Human Language Technology (HLT) conferences deserve special mention as they are a special joint conference between

Venue	Number of Papers	Years Active	Frequency
Journal	1291	1974-Present	Yearly
ACL	2037	1979-Present	Yearly
European ACL	596	1983-Present	2 Years
North American ACL	293	2000-Present	Yearly
Applied NLP	346	1983-2000	3 Years
COLING	2092	1965-Present	2 Years
HLT	957	1986-Present	2 Years
Workshops	2756	1990-Present	Yearly
TINLAP	128	1975-1987	Rarely
MUC	160	1991-1998	2 Years
IJCNLP	143	2005	—
Other	120	—	—

both speech- and text- practitioners, who ordinarily do not publish in the same venues.

Joseph and Radev [2007] have created the ACL Anthology Network, a citation graph over these papers. The experiments in this thesis rely on their citation graph and the bibliographic information they extracted. The ACL Anthology provides a text-only version of the documents, generated by PDFBox, which was used in the experiment after removing stopwords (common, content-less words) and non-word tokens. This preprocessed data is unfortunately a subset of all published papers: only 28 of the papers published in 2004 are available, though papers are available for 2005 and 2006. Figure 5.1 presents the number of papers for each year. All experiments were conducted on this data.

## 5.2 Computational Linguistics: The Hypotheses

In the first chapter, I introduced several questions that the techniques outlined in the preceding chapters can help me answer. In this section I enumerate some specific hypotheses about the history of ideas in Computational Linguistics and their adoption (or lack thereof) by researchers.

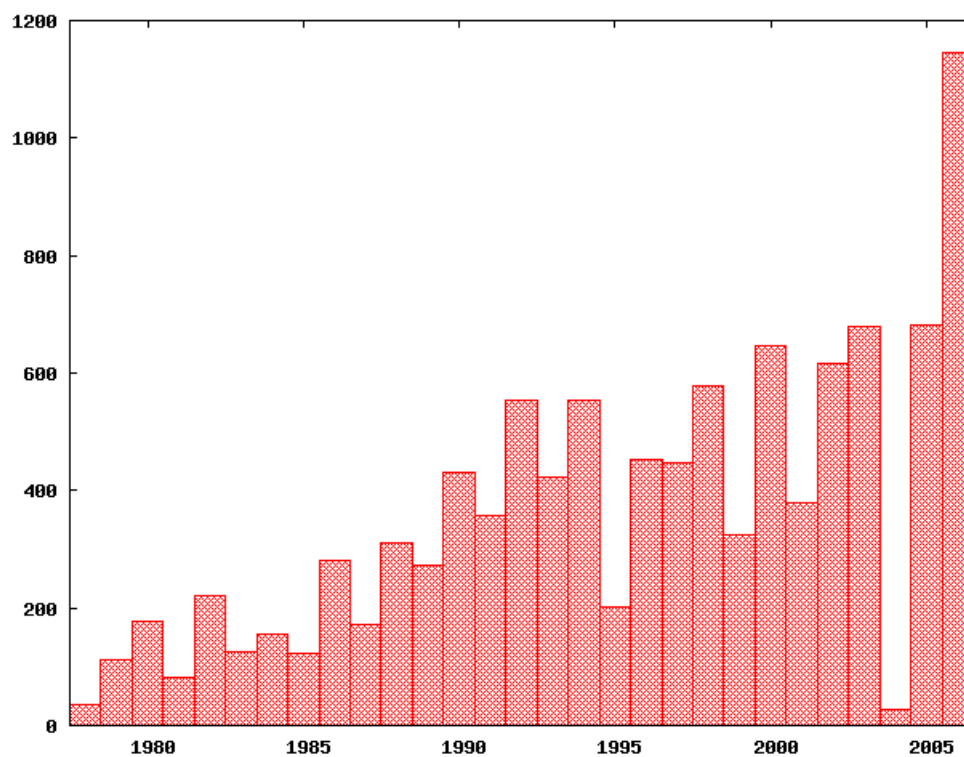


Figure 5.1: Number of papers published per year available in the ACL Anthology

### 5.2.1 The Birth of an Idea

Who is most involved in the creation of a new idea? Do established researchers who understand the field the best apply their experience to creating new ideas? Or do newer researchers seek to carve out a niche in what is new and different?

The introduction of statistics to NLP is a prime example of a new idea that has redefined a field. However, it is unclear who is most responsible for the adoption. Did new researchers who had not published in NLP bring new ideas to the field? Or did more experienced researchers identify a series of what Kuhn terms “anomalies” in the old paradigm that they could not solve and therefore overturned their own paradigm by establishing the use of probability?

Defining the “experienced” group to be those who published before 1990, and the “new” group to be those who did not publish before that year, I examine the probability that a researcher writing about probability before 1995 (and after 1990) was “new,” compared with the probability that they were old.

### 5.2.2 The Fate of Ideas

Most, if not all, ideas disappear, and as the progress of science increases, it seems inevitable that ideas rise and fall quicker than before. However, the researchers who are working on these ideas do not disappear, and they ought to continue publishing. That of course demands that I determine what ideas they are working on.

The case of Computational Linguistics provides a perfect candidate to examine the fate of an idea. Unification, as a grammatical formalism and approach to parsing, was one of the most influential frameworks in NLP throughout the 1980’s and into the early 1990’s. However, recently it has largely fallen from prominence. What became of the researchers? One might suppose that they moved, as a group, to a different specific topic. Or perhaps they dispersed more or less uniformly and are therefore indistinguishable from the larger community. Or, finally, they may have continued to publish as they had, but other, newer, researchers began to publish in other areas and by sheer numbers drowned out the researchers working on Unification. This last hypothesis would be consistent with the hypothesis from the previous section: experienced researchers continued to

work on what they worked on, while newer researchers were the ones instrumental in shaping the paradigms.

To choose between these hypotheses requires some amount of *post hoc* analysis. For example, I would like to determine what fields authors who were prominent in Unification moved to, as Unification has recently fallen out of favor as an approach. To determine this, I evaluate which authors were most associated with a topic before some date, and then what topics they were associated with after that date. Letting  $A_{d'}$  denote the set of authors in a given paper, I define:

$$\begin{aligned}
TF_y(k, j) &= p(z_{t>y} = k | z_{t<y} = j, t < y) \\
&= \sum_a p(z_{t>y} = k | a, y) p(a | z_{t<y} = j, t < y) \\
&= \sum_a p(z_{t>y} = k | a, y) \frac{p(z_{t<y} = j | a, y) p(a | t < y)}{p(z_{t<y} = j | y)} \\
&\propto \sum_a p(z_{t>y} = k | a, y) p(z_{t<y} = j | a, y) p(a | t < y) \\
&= \sum_a \left( \sum_d p(z_{t>y} = k | d) p(d | a, y) \right) \left( \sum_d p(z_{t<y} = j | d, y) p(d | a, y) \right) p(a | t < y) \\
&\propto \sum_a \left( \sum_{d:t_d > y, a \in A_d} \frac{C(z = k, d)}{N_d} \frac{N_d}{\sum_{d':t_{d'} > y, a \in A_{d'}} N_{d'}} \right) \left( \sum_{d:t_d < y, a \in A_d} \frac{C(z = j, d)}{N_d} \frac{N_d}{\sum_{d':t_{d'} < y, a \in A_{d'}} N_{d'}} \right) C(a) \\
&= \sum_a \left( \sum_{d:t_d > y, a \in A_d} \frac{C(z = k, d)}{\sum_{d':t_{d'} > y, a \in A_{d'}} N_{d'}} \right) \left( \sum_{d:t_d < y, a \in A_d} \frac{C(z = j, d)}{\sum_{d':t_{d'} < y, a \in A_{d'}} N_{d'}} \right) C(a, d, t_d < y)
\end{aligned} \tag{5.1}$$

, which is straightforward to calculate. Note that I make a few simplifying (but reasonable) assumptions. First,  $p(a | t < y)$ , the probability of an author's involvement before the epochal year, is proportional to the number of papers he wrote during in the period before that year. Second, I assume that  $p(d | a)$  is proportional to the number of words in the document, and simply assume that each author contributes to each document the same as the other authors.

To apply 5.1 to this problem, I consider the differences since 1995 between researchers who focused on Unification and the field as a whole. 1995 was the last year in which Unification was one of the most dominant topics in NLP and two years after the release of the Penn TreeBank, which is a corpus of hand-parsed sentences done outside of the unification frameworks. For comparison, I

will also analyze members of the Parsing subcommunity, which occupied a position not dissimilar to Unification, but has maintained more or less the same level of influence, even until the present time.

### 5.2.3 Paradigm Merger: Gazdar

An alternative explanation to the birth and death of paradigms that the two previous sections assumed is that two paradigms can merge. Gazdar [1996] wrote about a paradigm merger in Computational Linguistics between the LOGIC community and the NGRAM community. He points in particular to the task of parsing, claiming that ideas from the LOGIC group and the NGRAM group would fuse into some new idea. His assertion advances two hypotheses: first, there must have been both a LOGIC community and a NGRAM community already in existence, and second, that an actual merger of paradigms took place.

To test his hypotheses, I will focus on subcategories of NLP directly relevant to parsing. If there was in fact an NGRAM group, there should be some amount of work being done on statistics, probability, and machine learning techniques such as classification. And, if there was a merger, there should be substantial overlap between both LOGIC and NGRAM topics in parsing papers around the time of publication of his article.

## 5.3 Model Estimation

### 5.3.1 Collapsed Gibbs Sampling

Gibbs sampling is a method of estimating the parameters by repeatedly resampling variables given all the other variables until convergence is reached. This section presents derivations for a Gibbs sampler for the Shoulders of Giants model based in part on the derivations in Griffiths and Steyvers [2004].

After collapsing all the parameters, The likelihood of the data given the parameters is given

by:

$$p(\vec{w}, \vec{z}, \vec{c}, \vec{o} | \eta, \mu, \vec{\nu}, \alpha) = \int d\vec{\beta} p(\vec{\beta} | \eta) \prod_d \int d\vec{\psi}_d \int d\lambda_d \int d\vec{\gamma}_d \prod_i^{N_d} p(w_{di} | z_{di}, \vec{\beta}_{z_{di}}) \cdot p(z_{di} | c_{di}, o_{di}, \theta_{c_{di}}, \vec{\psi}_d) \cdot p(\vec{\psi}_d | \alpha) \cdot p(c_{di}, o_{di} | \lambda_d, \vec{\gamma}_d) \cdot p(\vec{\gamma}_d | \mu) \cdot p(\lambda_d | \vec{\nu})$$

The posterior for  $z_{di}$  is proportional to the product of the prior and the likelihood:

$$p(z_{di} | \vec{z}_{d-i}, \vec{z}_{-d}, w_{di}, c_{di}, o_{di}, \cdot) \propto \int p(z_{di} | c_{di}, o_{di}, \psi_d, \cdot) p(\psi_d | \alpha) d\psi_d \cdot \int p(w_{di} | z_{di}, \vec{\beta}_{z_{di}}, \cdot) p(\vec{\beta}_{z_{di}} | \eta) d\vec{\beta}_{z_{di}}$$

Letting  $o_{di} = 1$ —indicating that  $z_{di}$  is drawn from the original topics  $\psi_d$ —leaves for the particular cases:

$$p(z_{di} | \vec{z}_{d-i}, \vec{z}_{-d}, w_{di}, c_{di}, o_{di} = 1, \cdot) \propto \int p(z_{di} | o_{di} = 1, \psi_d, \cdot) p(\psi_d | \alpha) d\psi_d \cdot \int p(w_{di} | z_{di}, \vec{\beta}_{z_{di}}, \cdot) p(\vec{\beta}_{z_{di}} | \eta) d\vec{\beta}_{z_{di}}$$

$$p(z_{di} | \vec{z}_{d-i}, \vec{z}_{-d}, w_{di}, c_{di}, o_{di} = 0, \cdot) \propto p(z_{di} | o_{di} = 0, c_{di}, \theta_{c_{di}}, \cdot) \cdot \int p(w_{di} | z_{di}, \vec{\beta}_{z_{di}}, \cdot) p(\vec{\beta}_{z_{di}} | \eta) d\vec{\beta}_{z_{di}}$$

This first equation is analogous to the standard collapsed Gibbs updates for Latent Dirichlet Allocation presented in Griffiths and Steyvers [2004]. The second is analogous to not collapsing the variables. Omitting the details of the integrations:

$$p(z_{di} = k | \vec{z}_{d-i}, \vec{z}_{-d}, w_{di}, c_{di}, o_{di} = 1, \cdot) \propto \frac{C(\vec{z}_{d-i} = k, \vec{o}_{d-i} = 1) + \alpha}{C(\vec{o}_{d-i} = 1) + K\alpha} \frac{C(\vec{w}_{-di} = w_{di}, z_{-di} = k) + \eta}{C(\vec{z}_{-di} = k) + V \cdot \eta}$$

$$p(z_{di} = k | \vec{z}_{d-i}, \vec{z}_{-d}, w_{di}, c_{di}, o_{di} = 0, \cdot) \propto \theta_{dk} \cdot \frac{C(\vec{w}_{-di} = w_{di}, \vec{z}_{-di} = k) + \eta}{C(\vec{z}_{-di} = k) + V \cdot \eta}$$

, where  $C(\vec{z}_{d-i} = k, \vec{o}_{d-i} = 1)$  denotes the number of instances in document  $d$  besides the current one are both original and assigned topic  $k$ . The notation is expanded analogously for the other instances. Both of these equations have an intuitive explanation: the probability of assigning topic  $k$  to the  $i$ 'th word is proportional to the fraction of times that topic is used times the fraction of times that the word is generated from topic  $k$ 's vocabulary.

The posterior for  $o_{di}$  is given by:

$$p(o_{di} | \vec{o}_{d-i}, \vec{o}_{-d}, z_{di}, c_{di}, o_{di}, \cdot) \propto \int p(o_{di} | \lambda_d, \cdot) p(\lambda_d | \vec{v}, \cdot) d\lambda_d \cdot p(z_{di} | o_{di}, c_{di}, \cdot)$$

Again, splitting up the derivations for the two different assignments to  $o_{di}$  leaves:

$$\begin{aligned} p(o_{di} = 1 | \vec{o}_{d-i}, \vec{o}_{-d}, z_{di}, c_{di}, o_{di}, \cdot) &\propto \int p(o_{di} = 1 | \lambda_d, \cdot) p(\lambda_d | \vec{v}, \cdot) d\lambda_d \cdot \int p(z_{di} | o_{di} = 1, c_{di}, \psi_d) p(\psi_d | \alpha) d\psi_d \\ &\propto (C(o_{d-i} = 1) + \nu_1) \cdot (C(z_{d-i} = k, \vec{o}_{d-i} = 1) + \alpha) \\ p(o_{di} = 0 | \vec{o}_{d-i}, \vec{o}_{-d}, z_{di}, c_{di}, o_{di}, \cdot) &\propto \int p(o_{di} = 0 | \lambda_d, \cdot) p(\lambda_d | \vec{v}, \cdot) d\lambda_d \cdot p(z_{di} | o_{di} = 0, c_{di}, \cdot) \\ &\propto (C(\vec{o}_{d-i} = 0) + \nu_0) \cdot (\theta_{c_{di} z_{di}}) \end{aligned}$$

$c_{di}$  only needs to be resampled when  $o_{di} = 0$ , so:

$$\begin{aligned} p(c_{di} | \vec{c}_{d-i}, \vec{c}_{-d}, z_{di}, o_{di} = 0, \cdot) &\propto \int p(c_{di} | \gamma_d, \cdot) p(\gamma_d | \mu, \cdot) d\gamma_d \cdot p(z_{di} | c_{di}, \theta_{c_{di}}, o_{di} = 0) \\ &\propto (C(\vec{c}_{d-i} = c_{di}) + \mu) \cdot \theta_{c_{di} z_{di}} \end{aligned}$$

### 5.3.2 IdeaRank

IdeaRank is computed using the iterative algorithm proposed in Page et al. [1998]. The implementation is parallelized using a reduce-scatter operation analogous to MapReduce. [Dean and Ghemawat, 2008] Computation is rapid: 100 iterations takes approximately 5 minutes on a cluster of 5 dual-core machines.

# Chapter 6

## Results and Discussion

This chapter presents the results of applying the Shoulders of Giants Model and the other techniques outlined in the preceding chapters to the field of Computational Linguistics. I first give an overview of the trends and topics over the past thirty years, and then proceed to analyze the origin, progress, and participants in the Statistical Revolution.

### 6.1 Topics

This section gives an overview of the topics I found in Computational Linguistics. For initial analysis I extracted 60 topics from the ACL Anthology. Moreover, to get a sense of the topics' differing places in times, I found it useful to understand the importance of topics over the scope of the anthology. Note that unlike the Dynamic Topic Model or the Topics over Time model, this model has no explicit discussion of time. However, it is quite simple to derive an empirical estimation, using a simple calculation. I am interested in the empirical probability of a topic given the current

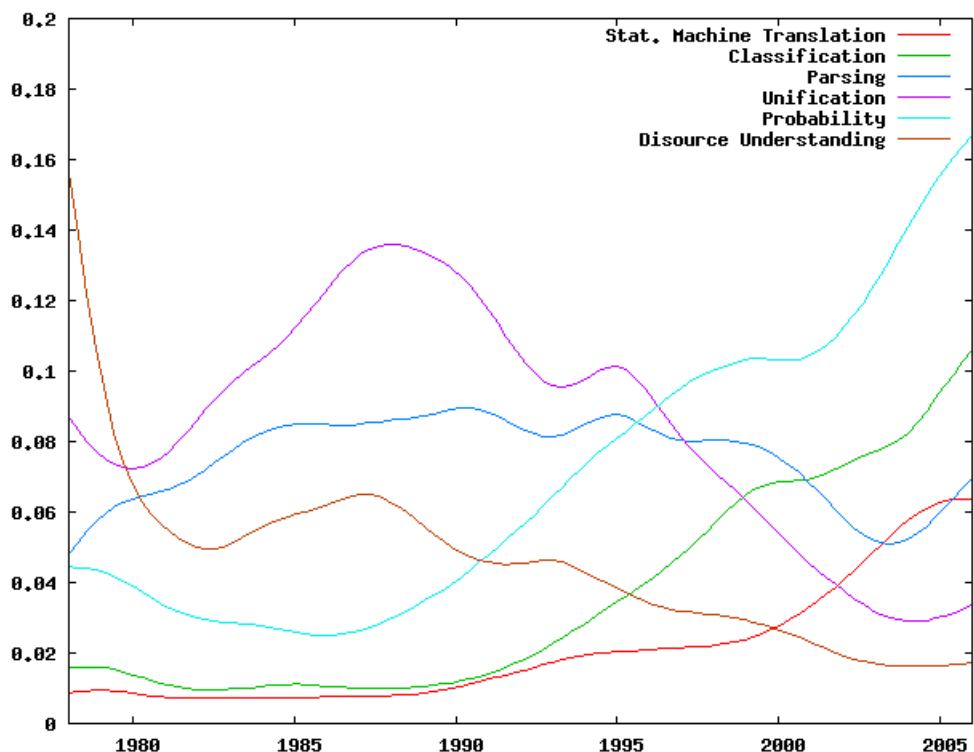


Figure 6.1: Trends: Prominence of Several Topics in Computational Linguistics

year:

$$\begin{aligned}
 \hat{p}(z|y) &= \sum_{d:t_d=y} \hat{p}(z|d)\hat{p}(d|y) \\
 &= \frac{1}{C} \sum_{d:t_d=y} \hat{p}(z|d) \\
 &= \frac{1}{C} \sum_{d:t_d=y} \theta_d
 \end{aligned}$$

, where I simply set  $\hat{p}(d|y)$  to the uniform distribution over documents published in that year. The result of this transformation for several of the topics is plotted using approximated cubic splines in Figure 6.1 make a basic trend graphs, while isolated graphs depicting several more topics are available in Figures 6.2, 6.3, and 6.4.

Several trends are immediately apparent. Initially, but for a strikingly short period of time,

discourse analysis was the predominant mode of reasoning, with almost one-quarter of the total topics in the first year plotted. Then, throughout the 1980's, Unification together with Parsing seemed dominant. Parsing has remained more or less constant in prominence, showing only a small decline in the past few years. Papers about unification, however, have fallen steeply since 1995, from a height of over 14% of the content of all papers down to under 4%. Simultaneously, concepts related to probability and machine learning have exploded since 1990, to the point where Probability, Classification-based Learning, and Machine Translation collectively account for almost one-third of all the topic mass in 2006.

## 6.2 Computational Linguistics: The Answers

This section discusses the forces and participants in the Statistical Revolution in Computational Linguistics. I present evidence to address the hypotheses discussed in the preceding chapters.

### 6.2.1 Paradigm Shift: The Rise of Probability

What started the Statistical Revolution? I calculated the probability that a paper written after 1990 and before 1995 was written by someone who had not published prior to 1990. Table 6.2.1 presents the proportions for select topics in for both groups, as well as the relative differences between them. One can immediately see the difference between the groups: new researchers are far more likely to take up new topics like Classification and Machine Translation than those who were publishing before. Older ideas like Unification were less likely to be worked on by newer people. This suggests that here it is new researchers—and not old ones—who are most likely to seize upon new ideas.

However, fully one-third of authors writing about probability even at this early period had published before, indicating that the ideas were not wholly from people outside the ACL. To explore this, I examined which of the experienced researchers most affiliated with Probability. Those writers are in Table 6.2.1. Referring back to their papers in the ACL Anthology, I discover that many of them published just once or twice prior to 1990, almost all in 1988 and 1989. These papers are:

1. [A88-1010] Grishman, Ralph, Chitrao, Mahesh V. Evaluation Of A Parallel Chart Parser

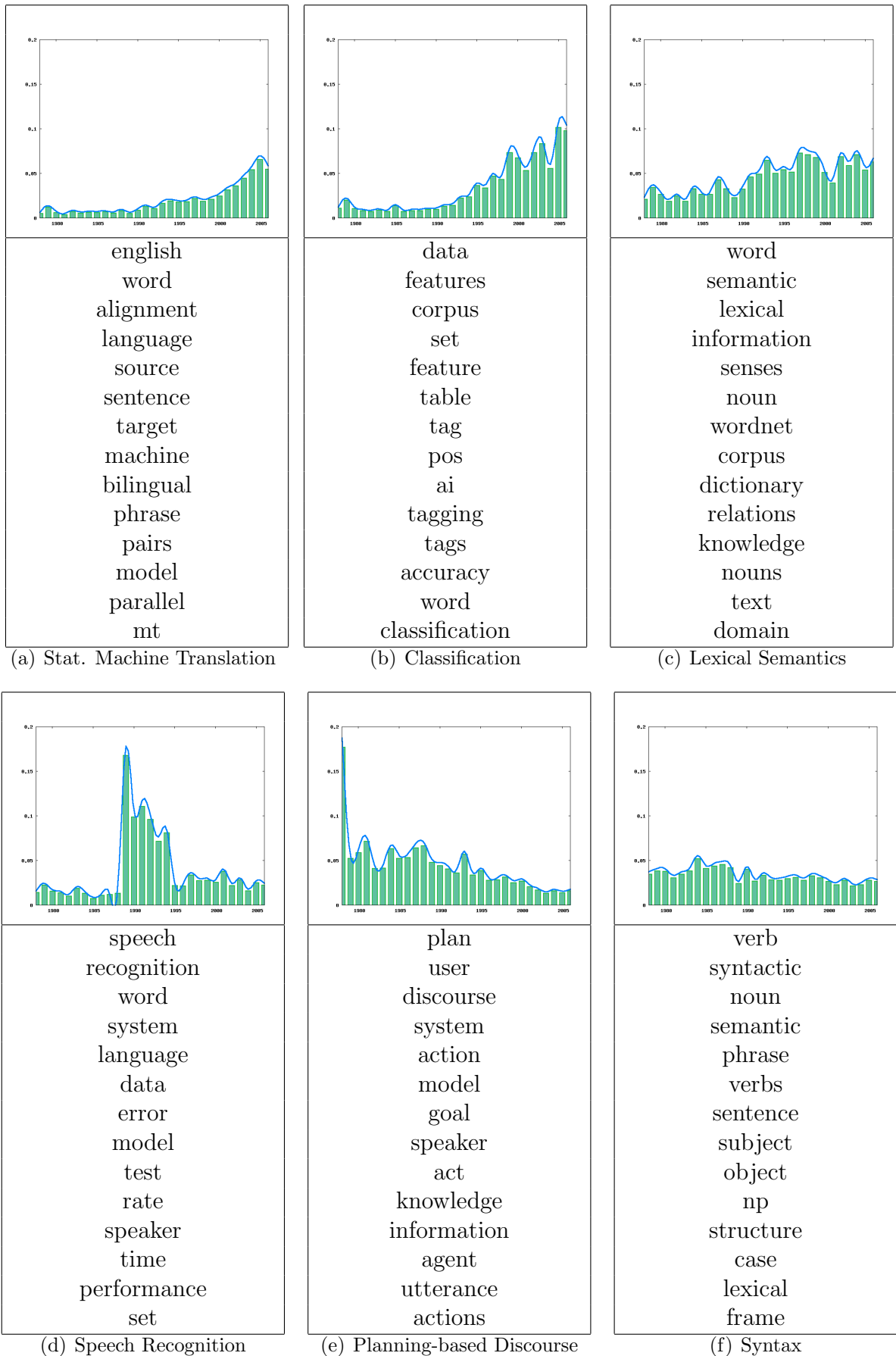


Figure 6.2: Vocabularies and Prominence for Topics at Different Times

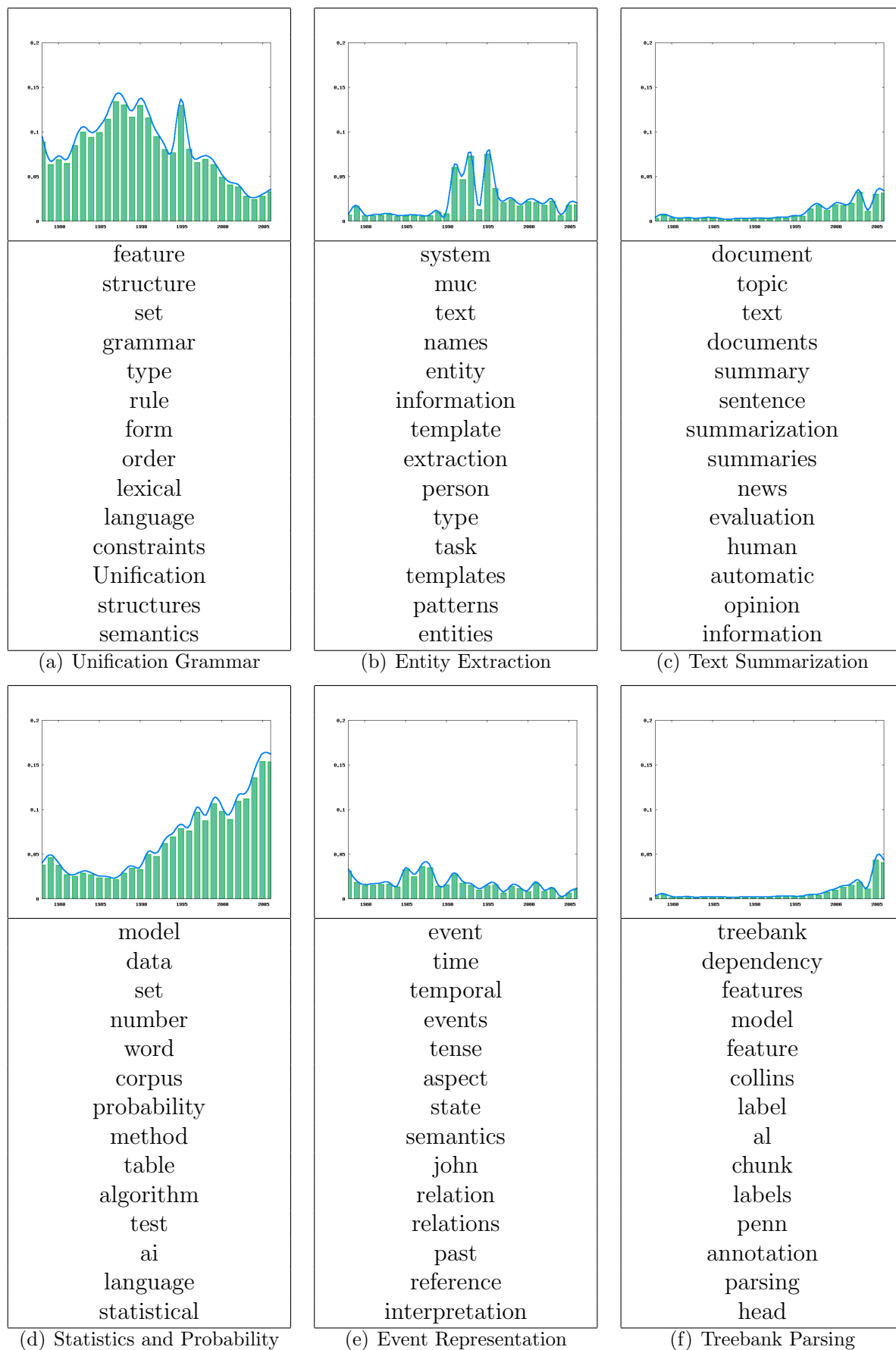


Figure 6.3: Vocabularies and Prominence for Topics at Different Times

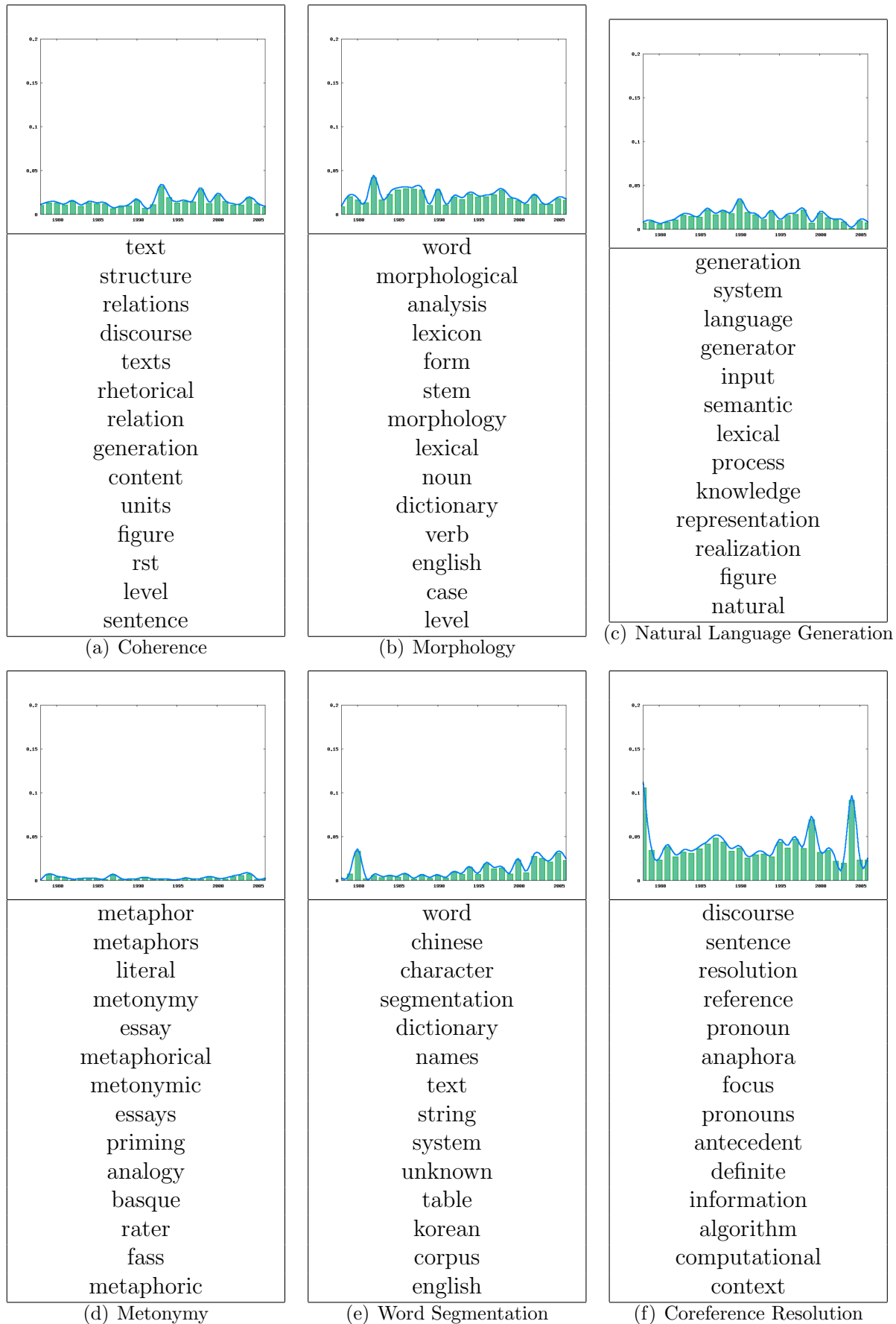


Figure 6.4: Vocabularies and Prominence for Topics at Different Times

Topic	New Researcher
Probability	0.663
Classification	0.624
Unification	0.425
Syntax	0.473
Parsing	0.518
Morphology	0.559
Language Generation	0.468
Information Retrieval	0.648
Speech Recognition	0.82
Total	0.522

Table 6.1: Probability that a paper about a topic written between 1990 and 1995 was written by someone new to the field

Author	Probability (1990-1995)
Matsunaga, Shoichi	0.392
Della Pietra, Vincent J.	0.335
Della Pietra, Vincent J.	0.327
Roukos, Salim	0.30
Iwayama, Makoto	0.305
Brown, Peter F.	0.303
Gale, William A.	0.298
Chitrao, Mahesh V.	0.284
Su, Keh-Yih	0.284
Yang, Yiming	0.268
Digalakis, Vassilios	0.262
Rohlicek, J. Robin	0.254
Church, Kenneth Ward	0.247
Tokunaga, Takenobu	0.238

Table 6.2: Loading for topic “Probability” during 1990-1995 for authors publishing prior to 1990

(ANLP, 1988)

2. [C88-1016] Brown, Peter F., Cocke, John, Della Pietra, Stephen A., Della Pietra, Vincent J., Jelinek, Frederick, Mercer, Robert L., Roossin, Paul S. A Statistical Approach To Language Translation (COLING, 1988)
3. [C88-1082] Matsunaga, Shoichi, Kohda, Masaki Linguistic Processing Using A Dependency Structure Grammar For Speech Recognition And Understanding (COLING, 1988)
4. [C88-2133] Su, Keh-Yih, Chang, Jing-Shin Semantic And Syntactic Aspects Of Score Function (COLING, 1988)
5. [C88-2136] Tokunaga, Takenobu, Iwayama, Makoto, Tanaka, Hozumi, Tadashi, Kamiwaki LangLAB: A Natural Language Analysis System (COLING, 1988)
6. [H89-1007] Roukos, Salim Integrating Speech And Natural Language. (Workshop On Speech And Natural Language, 1989)
7. [H89-1012] Boisen, Sean, Chow, Yen-Lu, Haas, Andrew R., Ingria, Robert J. P., Roukos, Salim, Stallard, David G. The BBN Spoken Language System. (Workshop On Speech And Natural Language, 1989)
8. [H89-2047] Digalakis, Vassilios, Ostendorf, Mari, Rohlicek, J. Robin Improvements In The Stochastic Segment Model For Phoneme Recognition (Workshop On Speech And Natural Language, 1989)
9. [H89-2062] Ostendorf, Mari, Rohlicek, J. Robin Segment-Based Acoustic Models With Multi-Level Search Algorithms For Continuous Speech Recognition (Workshop On Speech And Natural Language, 1989)

These 9 papers are the earliest publications for all but Kenneth Ward Church, who published several papers throughout the 1980's, but most famously [A88-1019] "A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text". The other papers share interesting common properties. First, they come from authors entirely outside the field. Almost all of these authors –with the exception of Grishman–had published nothing else in the ACL prior to these papers.

Moreover, all but one of these papers come just from two conferences, neither of which was the premier ACL conference. Many were papers from speech practitioners, marking a rare departure from Gazdar [1996]’s claim that speech and language practitioners rarely interchange ideas. In fact, the authors of [C88-1016] “A Statistical Approach to Machine Translation” are largely regarded as “speech people,” though they have no speech papers in the ACL. For instance, Peter Brown’s 1987 Doctoral Thesis is entitled “The Acoustic-Modeling Problem in Automatic Speech Recognition”.

## Interdisciplinarity: Speech and Text

That researchers in speech were among the first to bring statistics to the ACL—and that these papers by and large were not seen in the premier conferences—raises a small objection to Kuhn. In the Kuhn story, practitioners in one paradigm discover a series of anomalies in their paradigm: gaps that cannot easily be explained. Then, some particular scientist, the Newton of the time, develops a new paradigm that encompasses the anomalies in its own framework.

However, in the case of the Statistical Revolution, the paradigm shift came from outside the field, appearing on the outskirts. The new paradigm then had to work its way in, quickly finding itself vaulted to the central paradigm for doing research in NLP. Here, it took the outside injection of new ideas in the guise of an interdisciplinary conference along with researchers reaching outside their normal field of endeavor to revolutionize another. That is, in at least one instance, paradigm shifts have interdisciplinarity at their core.

### 6.2.2 The Decline of Unification

What became of the ideas dominant prior the Statistical Revolution? Figures 6.3 and 6.5 both demonstrate the decline in prominence of Unification in NLP. However, it was unclear what became of the authors: they still publish, but somehow Unification is still less important. To measure the effect, I evaluated  $TF_{1995}$  from 5.1 for Unification against all topics (for  $K=60$ ), and compared the importance of those topics for the authors after 1995 relative to the NLP community as a whole after 1995. 1995 was chosen because it was 2 years after the release of the Penn Treebank, the year before Gazdar [1996] was published, and the year before Unification began to fall out of favor.

Topic	Unification	Parsing	NLP	Rel. Difference
Probability	0.085	0.105	0.140	-39.3%
Classification	0.047	0.056	0.083	-43.4%
Treebank	0.014	0.017	0.022	-36.4%
Unification	0.137	0.089	0.060	+128%
Syntax	0.035	0.034	0.031	+12.9%
Parsing	0.114	0.159	0.081	+40.7%
Morphology	0.019	0.017	0.021	-9.52%
Machine Translation	0.022	0.024	0.042	-47.6%
Information Retrieval	0.012	0.014	0.021	-42.8%
Word Segmentation	0.009	0.014	0.019	-52.6%

Table 6.3: Differences in topic focus (after 1995) between authors focused on Unification prior to 1995 and the NLP community as a whole (K=60)

Table 6.2.2 presents topics that have substantially different proportions than the community as a whole, or are of interest separately.

Unsurprisingly, the authors who focused on Unification prior to 1995 continue to do so, despite the decline of Unification in the community at large. Interestingly, they do not focus on Syntax at a much higher rate than the community as a whole, but they do seem to seem largely to avoid the topics relating to probability than average. The Parsing comparison group seems to fall somewhere in the middle between the field as a whole and those focused on Unification. They are more likely to employ machine learning techniques like Classification and more likely to use probabilistic models in their papers.

Regardless, both groups were more reluctant to take up Probability and Classification than the field as a whole. This analysis further supports the Kuhn incommensurability hypothesis: members of the Unification community remained within Unification, while newer entrants adopted the newer ideas instead.

### 6.2.3 Paradigm Merger?

Gazdar [1996] predicted a paradigm merger between the LOGIC community and the NGRAM community. Instead, the results in the preceding chapter indicate that statistical approaches have merely been on the rise, and that LOGIC approaches have all but disappeared. Focusing more

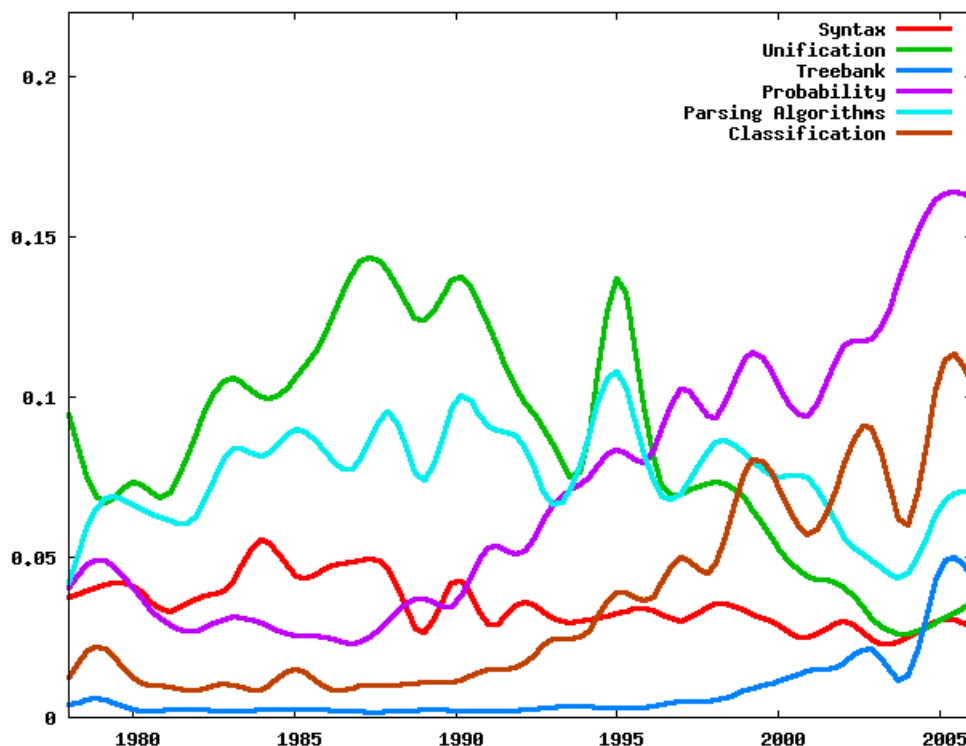


Figure 6.5: Relative Prominence of Topics Related to Parsing

sharply, Figure 6.5 highlights topics related to parsing, which has been clearly one of the most consistently influential areas in NLP. Prior to the 1990's, statistical and machine learning approaches (that is, the NGRAM community's tools) to parsing and tagging simply did not exist in any appreciable sense, while Unification—the dominant non-statistical approach—was clearly the most important framework. Thus, Gazdar's portrayal of 1980's NLP as two pre-existing communities is incorrect. Instead, there was simply one framework (Gazdar's LOGIC community) in the 1980's, and so the merger he describes is not possible. However, by the mid-1990's, Probability and Classification both were important topics in NLP: in Figure 6.5, Probability intersects with Unification in 1996, the same year that Gazdar's article was published. A modified version of Gazdar's hypothesis, then, is that the quick rise of Probability and Classification was quickly replaced by a “merger” between topics related to the LOGIC community and topics related to the (new) NGRAM community. Instead, Probability and Classification, unchanged, ascend quickly to prominence, while Unification, the core idea of the LOGIC community for parsing, quickly disappears.

### 6.3 IdeaRank

This section applies IdeaRank—the modified version of PageRank based on semantic analysis— Figure 6.6 lists the top papers by Citation Count, PageRank (as computed by Joseph and Radev [2007]) and IdeaRank for various numbers of topics. Focusing first on PageRank and Citation Count first: the top 10 are actually substantially different. In particular, “A Stochastic Parts Program and Noun Phrase Parser From Unrestricted Text” the #1 entry by PageRank is the fourth in Citation Count, while “Finding Clauses in Unrestricted Text By Finitary And Stochastic Methods,” #2 by PageRank, does not even show up in the top 10. In fact, it has been cited only 5 times (compared with 226 times for the #1 entry). This is in fact an artifact of PageRank: these two papers mutually cite each other, and the PageRank assigned to the higher paper ends up being shared with the otherwise not highly ranked paper.

Unfortunately, IdeaRank seems to suffer from the same problem, as the #4 entry for  $K=60$  is only cited by the #2 entry, though with  $K=100$  the paper is much lower at #30. In general, IdeaRank seems to be quite similar to the rankings provided by PageRank: 8 of the top 10 entries are the same, while only 4 of the top 10 papers by citation count are in either of the other rankings. However, IdeaRank seems reasonably stable even as  $K$  varies: both  $K=60$  and  $K=100$  overlap with 8/10 of the available papers. More qualitatively, the non-overlapping papers between the different rankings provide some insight. For instance, #8 by IdeaRank is cited by the #5 paper, which takes a substantial proportion ( $\approx 0.3$ ) of its topics from that paper.

	Citation Count
1	Building A Large Annotated Corpus Of English: The Penn Treebank
2	The Mathematics Of Statistical Machine Translation: Parameter Estimation
3	Attention Intentions And The Structure Of Discourse
4	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text
5	A Maximum-Entropy-Inspired Parser
6	A Maximum Entropy Approach To Natural Language Processing
7	Bleu: A Method For Automatic Evaluation Of Machine Translation
8	Three Generative Lexicalized Models For Statistical Parsing
9	A Maximum Entropy Model For Part-Of-Speech Tagging
10	Transformation-Based-Error-Driven Learning And Natural Language Processing
	PageRank
1	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text
2	Finding Clauses In Unrestricted Text By Finitary And Stochastic Methods
3	A Stochastic Approach To Parsing
4	A Statistical Approach To Machine Translation
5	The Contribution Of Parsing To Prosodic Phrasing In An Experimental Text-To-Speech System
6	Attention Intentions And The Structure Of Discourse
7	Building A Large Annotated Corpus Of English: The Penn Treebank
8	The Mathematics Of Statistical Machine Translation: Parameter Estimation
9	Deterministic Parsing Of Syntactic Non-Fluencies
10	The Semantics Of Grammar Formalisms Seen As Computer Languages
	IdeaRank (K=60)
1	A Statistical Approach To Machine Translation
2	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text
3	A Stochastic Approach To Parsing
4	Finding Clauses In Unrestricted Text By Finitary And Stochastic Methods
5	Attention Intentions And The Structure Of Discourse
6	The Mathematics Of Statistical Machine Translation: Parameter Estimation
7	Building A Large Annotated Corpus Of English: The Penn Treebank
8	Providing A Unified Account Of Definite Noun Phrases In Discourse
9	The Semantics Of Grammar Formalisms Seen As Computer Languages
10	Parsing as Deduction
	IdeaRank (K=100)
1	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text
2	A Statistical Approach To Machine Translation
3	The Mathematics Of Statistical Machine Translation: Parameter Estimation
4	Attention Intentions And The Structure Of Discourse
5	Building A Large Annotated Corpus Of English: The Penn Treebank
6	Focusing In Dialog
7	A Simple Rule-Based Part Of Speech Tagger
8	Parsing as Deduction
9	A Stochastic Approach To Parsing
10	Providing A Unified Account Of Definite Noun Phrases In Discourse

Figure 6.6: Top Papers in the ACL Anthology

# Chapter 7

## Conclusion

The findings in Computational Linguistics presented in this thesis largely support the Kuhn [1962] hypothesis: science progresses not by small iterative refinements, but by successive revolutions. At each revolution, not only the tools for addressing problems but also the kinds of problems that can be addressed are fundamentally different from those of the previous paradigm. The models and techniques developed in this thesis provide a framework for automatically extracting and determining the vocabularies that are used by these paradigms. Applying these techniques to Computational Linguistics uncovered novel results in the history of the field. In addition to confirming Kuhn's hypothesis that most researchers cannot readily change paradigms, I have found that new researchers from outside the field are often essential in starting a new paradigm, and that they are the first to take up a new paradigm once it arises. Moreover, these paradigms are fundamentally incommensurable: two different paradigms do not refer to the same kinds of tasks, and therefore they cannot "merge," as some have suggested.

### 7.1 The Role of Interdisciplinarity

This thesis has also demonstrated the role that interdisciplinarity has played in bringing about the Statistical Revolution. At least in this one case, introducing two different fields to each other seems to have had the most profound impact on at least one of them. Somehow, the notion of

interdisciplinarity does not figure in Kuhn [1962]’s own paradigm. However, the mixing of two normally separate disciplines ought to be a prime area for the creation of new scientific paradigms. Old ideas from one paradigm might address the anomalies in the research of another, which might then trigger a revolution.

And indeed, the story of the Statistical Revolution depends precisely on interdisciplinarity. It is clear that it was speech researchers, and not text researchers in NLP, who triggered the paradigm shift. (To an outside observer, it might still seem strange that speech and text should be so separate, but a glance at the trends graphs of the previous chapters should help confirm it.) This is not to say, of course, that interdisciplinarity is essential for paradigm shifts. Indeed, Kuhn’s own favorite story of the Copernican Revolution in astronomy was of course brought about by an astronomer. Indeed, disciplinary thinking seems to be quite capable of revolutionizing itself, but the special role of interdisciplinary research—as the initiator of revolutions—should not be understated.

At a higher level, this thesis has demonstrated that applying sophisticated computational models from natural language processing to the problems posed by the social sciences provides a powerful way to form and test new hypotheses as well as a framework for performing quantitative analyses that can concretely measure the influence of language and the influence that outside forces have on language. Moreover, the hypotheses created by social scientists provide a rich new field of ideas that computational linguists can and should challenge themselves to solve. By seeking to answer real questions that have a larger impact, NLP practitioners will no doubt discover new tools and ideas that can further advance the state of the art. Indeed, it is possible that the interdisciplinary involvement of these two groups of researchers will initiate a new scientific revolution within one or both of these fields.

## 7.2 Future Work

An immediate direction for future work is to apply the techniques developed in this thesis to a different domain with a longer history. The techniques developed in this thesis should be generally applicable to any domain, but an analysis of a field outside the author’s domain knowledge would be particularly interesting.

A different direction would be to study the small. The statistical models in this thesis were particularly well suited to extracting the vocabularies used to describe paradigms of research: macro-scale topics that cover broad swathes of research interests. Instead, one could envision models that emphasized small, nuanced vocabularies. Such models would be able to track the progress of “normal science,” and could be used to strengthen the results generated by the current implementation of IdeaRank.

Nevertheless, the work in this thesis has demonstrated that real problems in social science can be answered with greater precision and understanding by using techniques drawn from natural language processing. Of course, the work of theorists and human analysts is still crucial, but techniques like the ones derived here can help them to automate and articulate their findings in novel and insightful ways.

# Bibliography

David Blei and John D. Lafferty. Dynamic topic models. *ICML*, 2006.

David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

David Cohn and Thomas Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*, 2001.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008. ISSN 0001-0782.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 233–240. ACM, 2007.

Eric Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101:5220–5227, 2004.

Gerald Gazdar. *Paradigm merger in natural language processing*, pages 88–109. Cambridge University Press, 1996.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1: 5228–5235, April 2004.

- Thomas Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence*, Stockholm, 1999.
- Mark T. Joseph and Dragomir R. Radev. Citation analysis, centrality, and the acl anthology. Technical Report CSE-TR-535-07, University of Michigan. Department of Electrical Engineering and Computer Science, 2007.
- Martin Kay. Functional unification grammar: a formalism for machine translation. In *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pages 75–78. Association for Computational Linguistics, 1984.
- Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University Of Chicago Press, 1962.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval, Ch. 18, 19, 20 (optional)*. Cambridge University Press, 2008.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.
- Scott McDonald and Michael Ramscar. Testing the distributional hypothesis: The influence of context on judgments of semantic similarity. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 611–616, 2001.
- Robert K. Merton. The matthew effect in science, ii: Cumulative advantage and the symbolism of intellectual property. *Isis*, 79(4):606–623, 1988.
- Ramesh Nallapati and William Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. *International Conference for Weblogs and Social Media*, 2008.
- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
- Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. 2006.