

Microevolutionary Language Theory

by

Michael Lloyd Best

Bachelor of Science
University of California, Los Angeles
1989

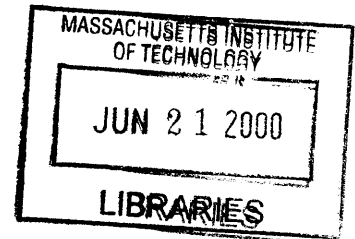
Master of Science
Massachusetts Institute of Technology
1996

ROTON

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY
at the
Massachusetts Institute of Technology
June, 2000

© Massachusetts Institute of Technology, 2000
All Rights Reserved



Signature of Author

Program in Media Arts and Sciences
March 1, 2000

Certified By

Pattie Maes
Associate Professor of Media Technology
Program in Media Arts and Sciences

Accepted By

Stephen A. Benton
Chair
Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Microevolutionary Language Theory

by

Michael Lloyd Best

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on March 1, 2000 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ABSTRACT

A new microevolutionary theory of complex design within language is proposed. Experiments were carried out that support the theory that complex functional design — adaptive complexity — accumulates due to the evolutionary algorithm at the simplest levels within human natural language. A large software system was developed which identifies and tracks evolutionary dynamics within text discourse. With this system hundreds of examples of activity suggesting evolutionary significance were distilled from a text collection of many millions of words.

Research contributions include: (1) An active replicator model of microevolutionary dynamics within natural language, (2) methods to distill active replicators offering evidence of evolutionary processes in action and at multiple linguistic levels (lexical, lexical co-occurrence, lexico-syntactic, and syntactic), (3) a demonstration that language evolution and organic evolution are both examples of a single over-arching evolutionary algorithm, (4) a set of tools to comparatively study language over time, and (5) methods to materially improve text retrieval.

Thesis Supervisor: Pattie Maes

Title: Associate Professor of Media Technology

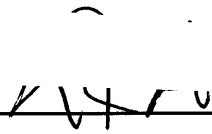
Microevolutionary Language Theory

by

Michael Lloyd Best

The following people served as readers for this thesis:

Reader



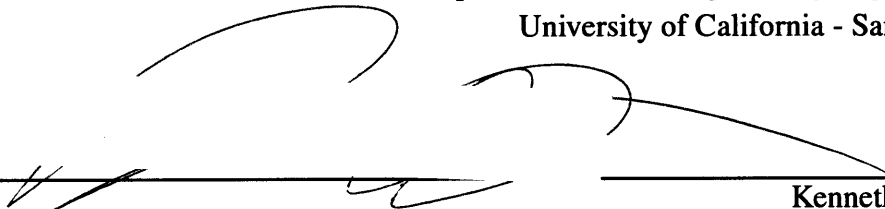
Richard K. Belew

Professor of Computer Science

Computer Science & Engineering Department

University of California - San Diego

Reader



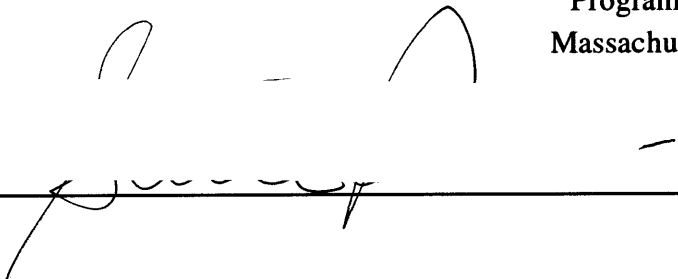
Kenneth Haase

Visiting Professor

Program in Media Arts and Sciences

Massachusetts Institute of Technology

Reader



Steven Pinker

Professor of Psychology

Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

Acknowledgments

First and foremost my thanks to my advisor Pattie Maes. She has supported me and my work with wisdom, clarity, and enthusiasm. I am humbled and thankful for her generosity.

To my committee I offer my sincere thanks; you have done me a great honor. My committee succeeded in bridging a range of subjects and provided significant insight and criticism. In addition, Ken Haase supervised my Masters thesis for which I am very grateful. Steve Pinker made time for (even encouraged) a number of meetings to thrash out issues and throw out ideas. Rik Belew provided a pinpoint critique on many problems with the dissertation. He also wins the prize for travelling the farthest, for that I am thankful.

I have been honored to know and work with an outstanding collection of colleagues over the years.

Warren Sack and Janet Cahn, members of the Machine Understanding Group, offered immeasurable help and insight. Warren and I have been talking about writing a paper together for the last five years; let's do it for real this time. Pushpinder Singh and Sara Elo also receive my thanks.

To the members of the Software Agents Group I am indebted — for practical matters such as machine maintenance and support, to bouncing ideas off of, to just providing a nice environment to hang out in. Thanks to Lenny Foner, Nelson Minar, Joanie Morris, Alex Moukas, Brad Rhodes, Sybil Shearin, Neil W. Van Dyke, David Wang, Alan Wexelblat, Jim Youll, and Giorgos Zacharia.

The News in the Future consortium has been a home for me and funded my research for quite a number of my years at the Lab. I am grateful to Walter Bender, Jack Driscoll, Felice Napolitano, and Rebecca Prendergast.

The Epistemology and Learning group has also been at times a home away from home for me. My thanks to the entire group and, most notably, Seymour Papert, Marina Umaschi Bers, David Cavallo, Claudia Urrea, and Jacqueline Karaaslanian.

A number of other Lab faculty have provided support, encouragement, and friendship. My thanks go to Justine Cassel, Judith Donath, Nicholas Negroponte, Sandy Pentland, Mitchel Resnick, Deb Roy, and Brian Smith.

The original idea, that some sort of language memetics might be realized through an analysis of internet discourse, is due to Richard Pocklington. He and I worked together on the very first experiments that provided a foundation for this dissertation. I am grateful for his enthusiasm to these ideas and generosity with them. Richard's advisor, Bill Durham, has been both an inspiration and source of ideas and insights for me. Bill served on my general exam committed for which I remain indebted.

Joey Berzowski and Beth Schneckeburger provided design and photo assistance (and friendship). Thanks go to them.

Linda Petterson steered me through the maze of Institute administrative requirements. Thanks Linda. Jon Ferguson, Will Glesnes, Dennis Irving, Greg Tucker, and Jane Wojcik have kept me with working computers, networks, and in an office for which I am grateful.

Finally, thanks and love to my family and in particular to my mother and father who copy edited a draft of this dissertation.

Table of Contents

CHAPTER 1	Introduction	13
1.1	Evolution as It Happens	14
1.2	Natural Traits	15
1.3	Replicators and Active Replicators	19
1.4	Roadmap to the Dissertation	22
CHAPTER 2	Language in Time and Space	23
2.1	Language Replicators	24
2.2	Summary	30
CHAPTER 3	The Chronica and CAMEL Software System	31
3.1	What Is a Chronica?	31
3.2	Globe Chronicon	32
3.3	Overview of NetNews	33
3.4	Netnews Chronica	36
3.5	Summary of Chronica	37
3.6	CAMEL System Overview	38
3.7	Stopwords and Stemming	39
3.7.1	Example	41
3.8	Vector Space Representation	42
3.8.1	Example	44
3.9	Text Clustering	45
3.10	Correlation Coefficients	47
3.11	Accounting for Statistical Artifact	48
3.11.1	Neutral shadow	52
3.11.2	Problems with timeseries	57
3.12	Chronica Revisited	60
3.13	Stylostatistical Features for Two Chronica	65
3.14	Summary	69
CHAPTER 4	Active Language Replicators	71
4.1	The Centrality of Level	71
4.2	Lexical Replicators	72
4.2.1	Globe active replicators	74
4.2.2	Summary and analysis	79
4.2.3	Clinton active replicators	81
4.3	Lexical Co-occurrence	83
4.3.1	Latent Semantic Indexing	83
4.3.2	Example	85

Table of Contents

4.3.3 Active lexical co-occurrence replicators	86
4.3.4 Summary and analysis	90
4.4 Lexico-syntax	91
4.4.1 English FDG	91
4.4.2 Noun phrase replicators	94
4.4.3 Active noun phrase replicators	97
4.4.4 Summary and evaluation	101
4.4.5 SVO replicators	102
4.5 Syntax	108
4.5.1 Evaluation	118
4.6 Summary	121
CHAPTER 5 Evolution as Algorithm	123
5.1 Campbell's Rule	124
5.2 The Lewontin-Campbell Computational Theory	126
5.3 Microevolutionary Language Corollary	127
5.3.1 Point #1 - individuals, traits, and heritability	127
5.3.2 Point #2 - variation	130
5.3.3 Point #3 - trait/fitness covariance	131
5.3.4 Quod Erat Demonstrandum	132
5.4 Hull-Dawkins Typological Theory	132
5.4.1 Hull-Dawkins meets language replicators	134
5.5 Units of Selection	137
5.5.1 Historical review	137
5.5.2 Cui bono? in text	139
5.5.3 Cui bono? as explanation	139
5.6 Microevolution and Complex Design	141
5.6.1 Adaptive value and adaptation	143
5.6.2 An historical demonstration of adaptation	144
5.6.3 Pejoration and selective forces	147
5.7 Selection as a Strong Force	149
5.7.1 Nonselectionist forces	151
5.7.2 Neutral models	154
5.7.3 Accumulation of usage	155
5.7.4 Summary	157
5.8 The Size of the Units of Selection	157
5.9 Macroevolutionary Consequences of Microevolution	160
5.10 Summary	163
CHAPTER 6 Ecologies of Text	165
6.1 Models for Interacting Populations	165
6.2 Special Test Chronicon	166
6.3 Timeseries Cross-correlation	166

6.4 Negative Cross-correlations: Competition versus Predator/Prey.	169
6.5 Competition and Niche Behavior	171
6.6 Competition	172
6.7 Summary	174
CHAPTER 7 Replicators and Text Retrieval	175
7.1 Review of Text Retrieval Experiments	175
7.2 Queries	178
7.3 Natural Language Text Retrieval	182
7.4 Results with Active Noun Phrase Replicators	185
7.5 Summary	188
CHAPTER 8 Related Work	191
8.1 Evolution and Social Behavior	191
8.1.1 Evolutionary culture theories	192
8.1.2 Evolutionary psychology	193
8.1.3 Social learning theory	194
8.1.4 Memetics	194
8.1.5 Controversies in cultural evolution	197
8.2 Language	197
8.2.1 Phylogenetic language evolution	198
8.2.2 Glossogenetic language evolution	198
8.2.3 Critiques of the phylogenetic/glossogenetic split	200
8.2.4 Computer and formal models	200
8.3 Text	202
8.3.1 Corpuslinguistics and text analysis	202
8.3.2 Text retrieval	203
8.4 Media Laboratory Work	204
8.5 Summary	205
CHAPTER 9 Conclusions	207
9.1 What is it good for?	208
9.2 How is it different?	208
9.3 Where can we go from here?	210
9.3.1 New work in support of the Microevolutionary Language Theory	210
9.3.2 Improvements to the text analysis system	211
9.3.3 New studies and environments	212
9.4 Summary	212
Appendix A	215
References	219

Table of Contents

This dissertation proposes a new theory of complex design within language. It purports to answer the question:

Does complex functional design — adaptive complexity — accumulate due to the evolutionary algorithm at the simplest levels within human natural language?

In this dissertation I develop the thesis by (1) proposing a model in which appropriate microevolutionary dynamics within natural language can be tracked, (2) developing a large software system which identifies these dynamics over time within collections of text, (3) using this software system to amass evidence of evolutionary significant activity across collections of text composed of many millions of words, and (4) arguing that these results, when placed within a wider evolutionary theoretic framework, demonstrate complex design at a simple level. It is this hypothesis that I call the *Microevolutionary Language Theory*, to wit, complex functional design accumulates at the simplest levels (e.g., words, phrases) within natural language due to the process of evolution.

In developing this theory I offer a series of research contributions which include (1) an active replicator model of microevolutionary dynamics within natural language, (2) methods to distill evidence of evolutionary significant replicators, and a demonstration of evolutionary processes in action at multiple linguistic levels (lexical, lexical co-occurrence, lexico-syntactic, and syntactic), (3) a demonstration of Campbell's Rule within natural language: that language evolution and organic evo-

lution are both examples of a single over-arching evolutionary algorithm, (4) a set of tools to compare language across text, time, media, and community and (5) methods employing these techniques to materially improve text retrieval. Furthermore, through this work a number of other research results are touched on: (1) a comparative corpuslinguistic study of UseNet News discussions versus traditional print newspaper corpora, (2) a method to track “what’s hot” and “what’s not” within the news media, internet discussions, and similar media, (3) theoretical speculations on how the Microevolutionary Language Theory links to the traditional research program of historical linguistics, (4) methods to study competitive ecological interactions between text populations, and (5) an operationalization of the “meme” meme.

Central to these findings, and to the Microevolutionary Language Theory overall, is the concept of an evolutionarily *active replicator*. The rest of this chapter introduces this concept through a discussion of the most famous of Darwinian Poster Children: finches of the Galápagos archipelago. After a look into the lives of these finches, I will conclude this short chapter with a roadmap to the rest of the dissertation.

1.1 Evolution as It Happens

We see nothing of these slow changes in progress, until the hand of time has marked the lapse of ages, and then so imperfect is our view into long-past geological ages that we see only that the forms of life are now different from what they formerly were. (Darwin, 1964/1859, p. 84)

Charles Darwin was a doubter. He did not think it was possible to observe evolution and natural selection in the flesh — evolution as it happens. However, Darwin did think that evolution was recoverable from the frozen fossil record. And one of his powerful observations was that the evolutionary process could be read through a comparative analysis from the natural palimpsest. But in his view, evolutionary dynamics in progress were outside our means of observation (Darwin, 1964/1859). Happily, Darwin was wrong and today we have a group of empiricists observing organic evolution in the moment (e.g., quite recently Lenormand, Bourguet, Guillemaud & Raymond (1999) and many others).

Many researchers in the evolution of *language* have taken the same tack as Darwin, ignoring evolution in the moment and instead looking for evidence of language evolution either in the hoary monuments of linguistic history (the frozen fossils of language) or through a comparative analysis of contemporary languages synchron-

icly frozen in time and space (the worn palimpsest of language). Through these systems of study, language evolution becomes a pretty static business: language is frozen in the linguistic fossils of our past, in a synchronic analysis of our present, or in the Pleistocene with our human *bauplan*.

The Microevolutionary Language Theory develops a contrasting model in which the dynamics of language evolution are observable in the moment; language evolution returns to its active nature. The observation of evolutionary dynamics as they happen centers on the identification of appropriate units of language replication and selection. This provides the fundamental activity in building an active language replicator model. The same has been true for organic evolution: The evolutionary process has been observed in action, thanks to the careful and close attention to appropriate traits undergoing replication and differential selection.

Certainly, the most famous example of observing evolution in the flesh is the long-term observation of Darwin's finches on the Galápagos islands conducted by Peter and Rosemary Grant and their colleagues and popularized in Jonathan Weiner's *The Beak of the Finch* (1995). I turn to their work to illustrate the concepts of trait, replication, and active replication, by considering a particular finch (*Geospiza fortis*), a particular island of the Galápagos (Daphne Major), and a particular event (the drought of 1977).

1.2 Natural Traits

We behold the face of nature bright with gladness, we often see superabundance of food; we do not see, or we forget, that the birds which are idly singing round us mostly live on insects or seeds, and are thus constantly destroying life; or we forget how largely these songsters, or their eggs, or their nestlings are destroyed by birds and beasts of prey; we do not always bear in mind, that though food may be now superabundant, it is not so at all seasons of each recurring year. (Darwin, 1964/1859, p. 62)

Meet the medium ground finch (*Geospiza fortis*); a male specimen is pictured in Figure 1. This *G. fortis* was photographed on the Isla Santa Cruz, though this species has been identified on 13 of the 17 major islands which make up the Galápagos group (Grant, 1986). For this finch, life on the Galápagos can often be a drunken, heady affair. When rain levels are high, abundant foodstuffs lead to a relatively

carefree existence, not too dissimilar from the picture painted by Darwin: a life “bright with gladness.”



FIGURE 1. Male medium ground finch (*Geospiza fortis*) from Isla Santa Cruz of the Galápagos archipelago. From Feldman (1998).

But, in fact, there is significant yearly and seasonal variation in rainfall on these islands. These ups and downs in rainfall produce dramatic variation in the abundance of seeds and other foodstuffs relied upon by *G. fortis* (Grant, 1985). The annual rain fluctuations on Daphne Major from 1976 - 1982 are shown in Figure 2. One thing should be clear from this graphic, 1977 was a very dry year. And as

might be expected, that year proved to be anything but “bright with gladness” for the *G. fortis* population of Daphne Major.

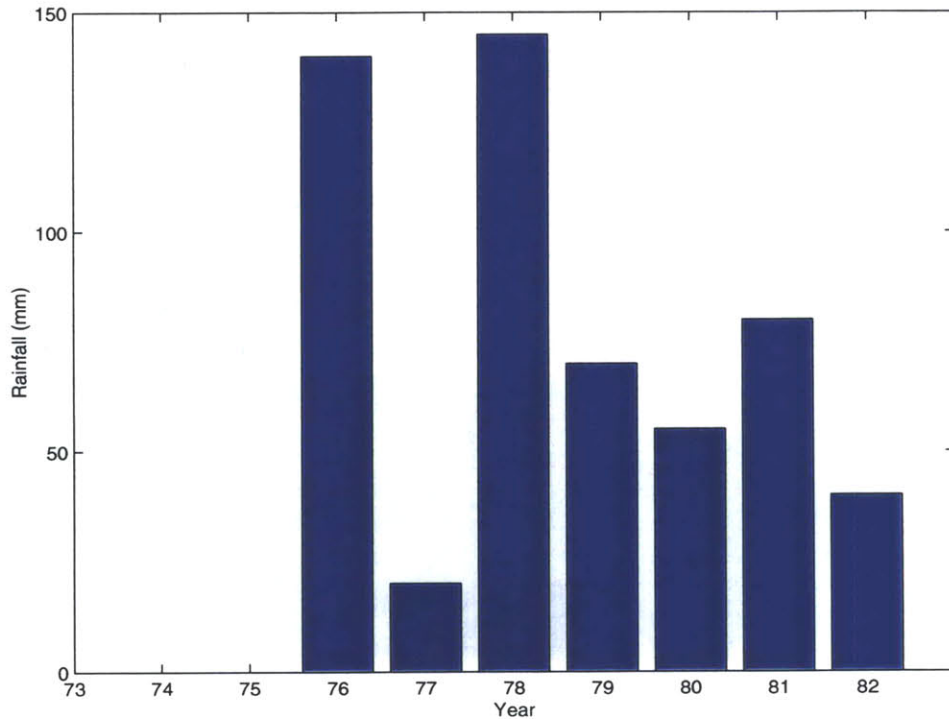


FIGURE 2. Rainfall measured on Daphne Major, 1976-1982. Drought of 1977 is clear. From Grant (1985).

When rainfall is heavy seeds are abundant, including an ample supply of the seeds most favored by *G. fortis*, small and soft seeds easily cracked and consumed such as the pistachio nut or *Heliotropium*. But in times of low rainfall, and in particular during the drought of 1977, the finches must contend with big and tough seeds, such as *Palo Santo* or seeds of the cactus *Tribulus*, which are difficult to eat (Weiner, 1995). In the actuarial book-keeping that describes the struggle for life, these seeds take more energy to harvest and consume per energy delivered than the favored soft and small seeds. However, the difficulty experienced when consuming these big, tough seeds is not uniform across all *G. fortis*. As it happens, those finches with bigger beaks, especially deeper beaks, have an easier time than those with smaller beaks (Grant, 1986). Figure 3 shows the variation in bill depth among

the medium ground finches on Daphne Major during times of usual rainfall. And Figure 4 depicts the cross-section of beak measured as “bill depth.”

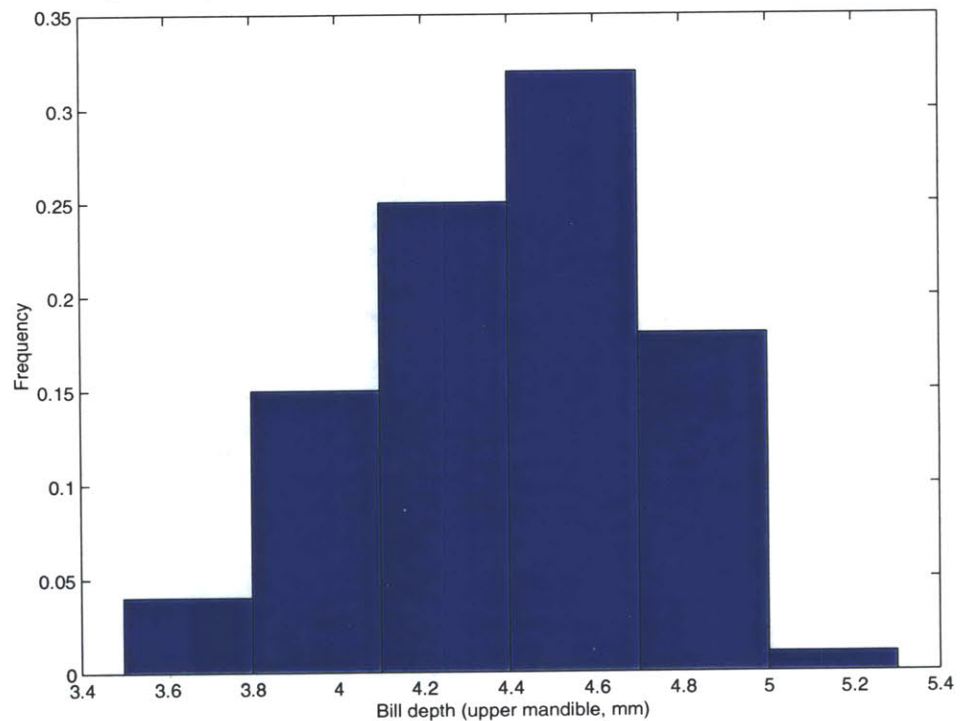


FIGURE 3. Frequency distribution of bill depth (upper mandible) of adult male population of medium ground finch (*Geospiza fortis*) on Daphne Major ($n = 89$). From Grant (1986).

During the drought of 1977, seed, beak, and rain all came together quite dramatically. As days turned into weeks free from usual rainfall the available seed biomass crashed to near zero. Finches had to forage for and feed off the hated tough and big seeds to survive. The outcome was predictable: as the desirable seed availability

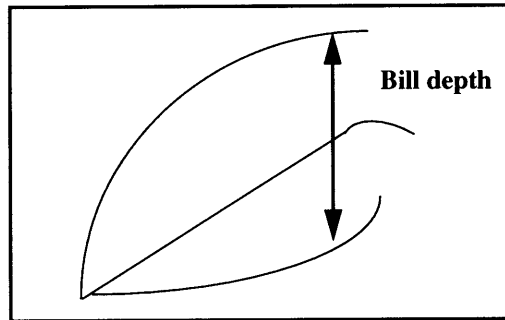


FIGURE 4. The cross-section measurement for bill depth (upper and lower mandible).

crashed so did the population of *G. fortis* (see Figure 5). The few finches that did survive were those lucky enough to have the deep beaks most suitable for cracking and feeding on tough and big seeds (Boag & Grant, 1984; Grant, 1985).

1.3 Replicators and Active Replicators

This is a story, to be sure, of “Tho’ Nature, red in tooth and claw.” But it also serves to illustrate the fundamental evolutionary concepts I will be identifying within natural language: replicators and active replicators. The bill depth trait of the medium ground finch is a property of the morphology of these birds that is the expression, primarily, of inherited genetic coding. This genetic, heritable link has been directly established by Grant (1986) and his colleagues. The genetic code and, if we allow ourselves a bit of terminological play, the trait itself, are *replicators*, insofar as they are bits of information that reoccur over time. Indeed the evolutionary theorist Richard Dawkins describes a replicator as simply “anything in the universe of which copies are made” (Dawkins, 1982, p. 83). But within the evolutionary process not all replicators are created equal. Special attention is given to *active replicators*, these are reoccurring entities that autocatalyze their own subsequent reoccurrence. In other words, their presence affects their chance of reappearing. As Dawkins puts it, an “*active replicator* is any replicator whose nature has some influence over its probability of being copied” (Dawkins, 1982, p. 83, emphasis in original). To take a page from Gregory Bateson, an active replicator is “a difference that makes a difference in its own chance of making a difference.”

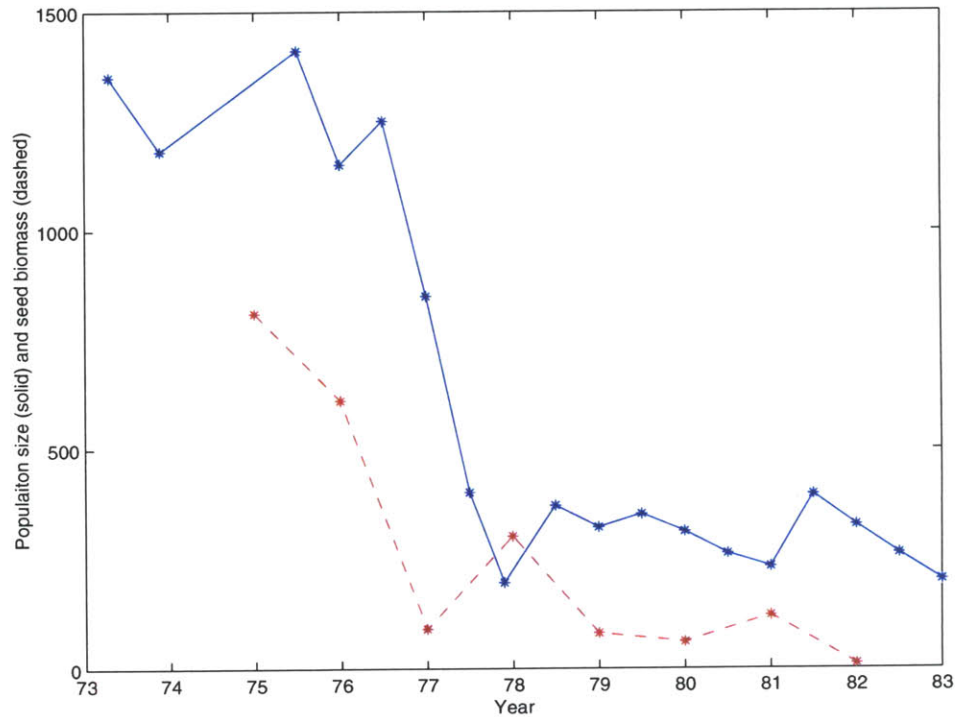


FIGURE 5. Population size of *Geospiza fortis* (solid) and small seed biomass (dashed) on Daphne Major from 1973 - 1982. Crash associated with drought of 1977 is apparent. From Grant (1985).

The beak depth trait of a *G. fortis* is a replicator. But is it an active replicator? The simplest way to answer this is to measure directly its correlation with survivability. In Figure 6, I show the fitness function for beak depth around the time of the drought as computed by Boag and Grant (1984). It demonstrates quite dramatically that the depth of the bill directly correlates with the chance of survival of the finch. Since beak depth is an inherited genetic trait, a finch with a deeper beak is more likely to survive, and in turn more likely to pass on this trait: The trait aids survivability, survivability increases reproductive success, and this in turn passes on the

trait. Thus, bill depth is an active replicator as its expression autocatalyses its reappearance in subsequent generations of finches.

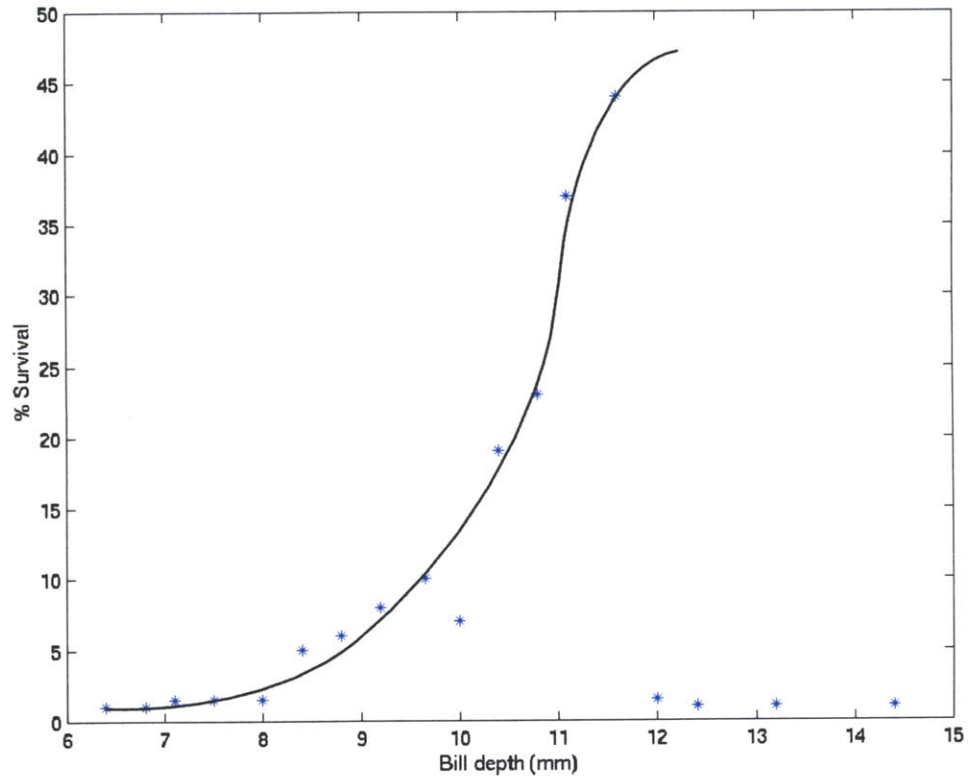


FIGURE 6. Fitness function with respect to bill depth (upper and lower mandible) for *Geospiza fortis* on Daphne Major from 1976 to 1978. Fitted line is hand drawn. From Boag and Grant (1984).

Active replicators sit at the centre of much of contemporary evolutionary theory (Williams, 1966; Dawkins, 1976). Steven Pinker has argued that this focus has been a “stunning success. It has asked, and is finding answers to, the deepest questions about life, such as how life arose, why there are cells, why there are bodies, why there is sex, how the genome is structured, why animals interact socially, and why there is communication” (Pinker, 1997, p. 43).

The Microevolutionary Language Theory explores complex adaptive traits at the simplest level within human natural language. And it does so by building evidence for active replicators within text collections in a manner not unlike that applied to these finches of the Galápagos. Concisely put, active language replicators are demonstrated empirically by:

- identifying linguistic traits that reoccur over time within a set of texts,
- arguing that relevant subsets of these texts describe an evolutionary lineage because they share copied traits,
- demonstrating that the appearance of some of these reoccurring traits correlates with a measure of survivability for these texts.

It turns out that many elements within human natural language have this autocatalytic, active property — from individual words to collocations, lexico-syntactic phrases, and perhaps even syntactic patterns.

This replicator-eyed approach to language, which is the central feature of the Microevolutionary Language Theory, provides a powerful conceptual integration of contemporary evolutionary theory with corpuslinguistic models of language use.

1.4 Roadmap to the Dissertation

This chapter has illustrated the fundamental concept of an active replicator by considering the beak of Darwin's finches. In the next chapter I will introduce the same sort of autocatalytic trait replicating instead within collections of text. Chapter 3 gives a fairly complete overview to the text analysis system, named CAMEL, that I have developed. Chapter 4 gives an overview of the principal results; I describe active replicators at multiple levels in language. In Chapter 5, I develop how these results are framed by and impact evolutionary and language theory. Chapters 6 and 7 serve to illustrate the sort of empirical studies and practical outcomes that emerge from the Microevolutionary Language Theory. In chapter 6 I trace competitive ecological interactions between populations of texts; Chapter 7 shows that attention to active replication can materially improve the precision of text retrieval engines.

This dissertation straddles three major fields of study: corpuslinguistics, evolutionary theory, and information retrieval. In Chapter 8, I review work across these disciplines relative to this dissertation, and relate the work to the concept of the "meme." And in Chapter 9, I end with my conclusions and a glance at some opportunities for future work.

Language in Time and Space

In Chapter 1 I told a story of microevolutionary dynamics of Darwin's finches. I showed how under heavy selection the beaks of these finches described active replicators; they autocatalysed (one way or the other) their reappearance in subsequent generations.

A similar story will be told in this chapter. I will show how lexico-syntactic traits, in this case two different noun phrases, act as active replicators over time. I will describe briefly the computational techniques used to distill these features, and will then track their rise and fall. As these noun phrases gain prominence within a given population of texts at a given time, a greater number of subsequent texts will be published within this same lineage. As these lineages share traits relative to the population as a whole, this is just the sort of autocatalytic process we are after.

The story of these active language replicators will serve to illustrate my general computational approach, and the type of results one can develop with it. In subsequent chapters (in particular Chapter 3), I will describe in far greater detail the collections of text analyzed and the software system used. Chapter 4 will give a full review of replicators at a variety of linguistic levels — lexical, lexical co-occurrence, lexico-syntactic, and syntactic.

2.1 Language Replicators

The bulk of results in this dissertation comes from the analysis of two different temporal corpora. These are collections of texts with a clear arrow of time running through them which I call *chronica*, to suggest the importance time plays (see Section 3.1). One of these collections is the “Globe” *chronicon* (in the singular) which is composed of all national and international news articles published in the print version of *The Boston Globe* from March 1, 1997 to December 16, 1998. It comprises eight million words of text. Another collection, the “Clinton” *chronicon*, is made up of all posts to the Usenet News (NetNews) newsgroup alt.politics.clinton from January 29, 1999 to March 5, 1999. It is composed of 6 million words of text. (The next chapter gives considerable details on these collections.)

The first step in analyzing these *chronica* is to cluster each of the texts around its principal topic area. For instance, the Globe collection may cluster on topics such as the war in Iraq or Social Security reform. The Clinton collection produces clusters that are less clear-cut usually; examples include discussions on congressional power and a thread dealing with certain legal challenges to the Clinton presidency. These clusters describe lineages within the *chronica*. Clustering is accomplished through traditional text analysis techniques: I compute the frequencies of particularly salient terms within the texts and then group them based on the degree to which they share these terms. If we graph over time the number of texts within a particular cluster we see the relative attention paid to some particular topical area against time. For instance, Figure 7 shows the timeseries for the 934 texts assigned to a particular cluster within the Globe *chronicon*. This cluster of texts deals with the scandal, during this time period, which centered on the relationship between White House intern Monica S. Lewinsky and President Bill Clinton (I call this the “Clinton/Lewinsky” cluster). Each point on this graph represents the number of texts published on this topic in a one-week interval. Clearly, there is a significant

boost in reporting midway through the date range. The week beginning February 2, 1998 was the most active with 31 articles published.

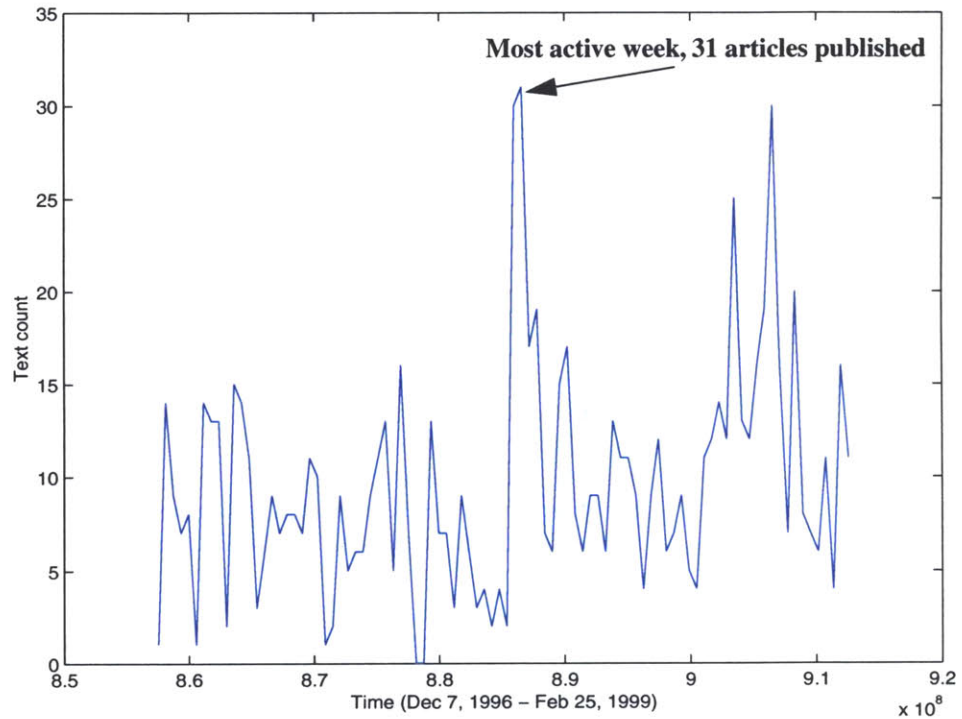


FIGURE 7. The 934 texts from the *Globe chronicon* assigned to a single cluster dealing with the scandal involving Monica S. Lewinsky and President Bill Clinton. Texts are bucketed at one-week intervals and time is represented in Unix format. Thus, each point represents the number of texts published in a given week on the topic.

Next I track the relative presence of noun phrases through the corpus. The first step is to identify such replicators by tagging and parsing the texts and extracting noun phrases. This process results in a list of noun phrases and how often each is used in any given text. For instance, the single noun phrase that was most frequent across the entire *Globe chronicon* was the proper noun “World War II,” the runner up was “White House official.” These noun phrases occurred 495 and 276 times respectively in the two years of texts studied (see Table 19 in Chapter 4).

Both of these analyses, the cluster timeseries and the noun phrase counts, tell us things about this collection of texts and its content over time. We know, for instance, when the Clinton/Lewinsky scandal heated up on the pages of *The Boston Globe* and we know what names appeared often in the texts.

However, I am not only interested in the total number of appearances for each noun phrase across the entire corpus. I also wish to know the relative presence of these lexico-syntactic replicators within each individual text — over time the rise and fall of relative usage. I have computed the number of times the phrases appear in each text and normalized for the text length. This normalized value indicates the percentage of any given text devoted to each particular noun phrase. (If I did not perform this normalization step, the analysis would be overly sensitive to variations in the length of articles.) Each replicator's measure of relative presence describes a trait for each text; it is important to understand why. Consider the “Monica S. Lewinsky” string. In many texts this noun phrase will not occur at all and in those texts this trait will be assigned a value of 0.0. But in some texts this trait scores as high a normalized value as 0.667. This does not mean that two-thirds of the text is made up of the string “Monica S. Lewinsky,” since our analysis is complicated by a number of factors (as will be explained in Chapter 3). However, it does mean that this trait is significantly expressed within the text relative to other traits with smaller values.

Note that the measurement of the normalized appearance of these phrases within the texts over time is *not* materially different from the measurement of beak depth of finches over time. They both describe expressed metric (real-valued) traits for these individuals.

In Figure 8, I have plotted the relative presence of the “Monica S. Lewinsky” noun phrase for those texts assigned to the Clinton/Lewinsky cluster. The timeseries of Figure 7 is repeated. Each point on the graph for the Lewinsky trait represents the average normalized appearance of the trait for that week. In other words, I average the relative presence of the noun phrase across all texts published for that week. The result is a metric trait normalized for both length of texts and number of texts. These values lie between zero and one. So in Figure 8 the graph has been scaled so that it is easily viewable when plotted along with the cluster count. To summarize Figure 8: The solid graph plots the week-by-week number of texts published on the Clinton/Lewinsky scandal in the *Globe* chronicon. The dashed graph plots the relative presence of the “Monica S. Lewinsky” noun phrase within these texts, week-by-week and normalized for length of text and number of texts. To a first approximation, it represents the percentage of text for the week occupied by the “Monica S.

Lewinsky” string. The higher the value, the more print is spent on that trait relative to print spent on other strings within this cluster.

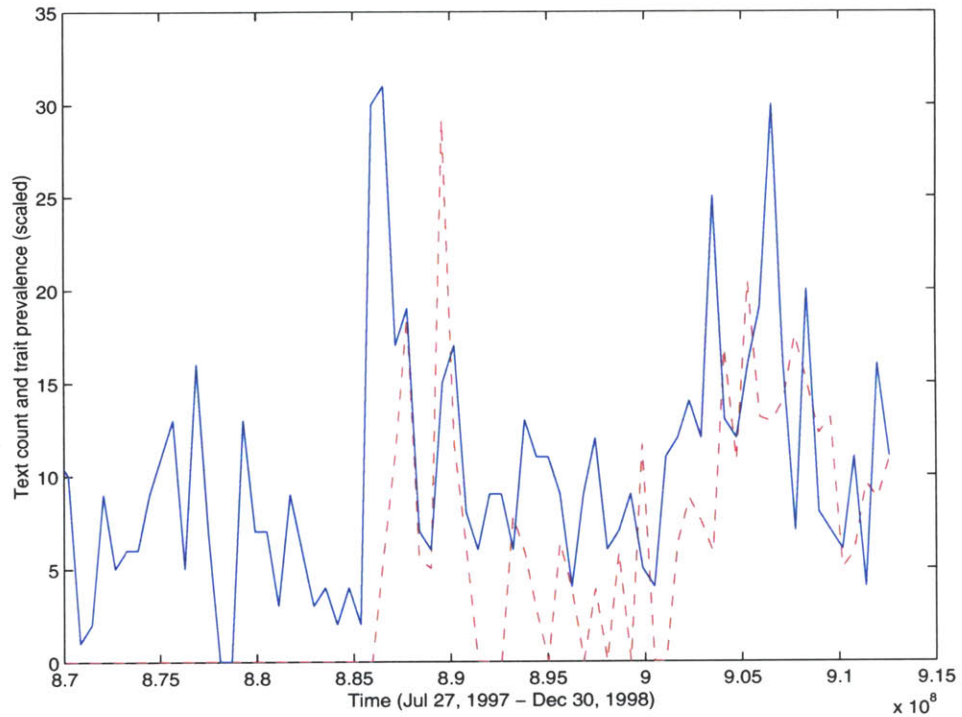


FIGURE 8. “Monica S. Lewinsky” trait (scaled, dashed) and number of articles printed on the Clinton/Lewinsky scandal within *The Boston Globe* ($r = 0.44, p < 0.00001$).

It should be apparent that a positive correlation exists between these two graphs. In fact the Pearson product-moment correlation is $r = 0.44, p < 0.00001$ (see Section 3.10). As more print on average is spent on the “Monica S. Lewinsky” string within a single text, more articles are published to the Clinton/Lewinsky cluster. And this is exactly the sort of autocatalytic process required for an active replicator.

In Chapter 1 it was noted that the deeper the finch’s beak the higher the survivability, and thus the more progeny, on average, that finch would have. And those children in turn would have, on average, deeper beaks. Bill depth was an expression of

an active replicator. Just so, as the “Monica S. Lewinsky” trait rises in prominence, more articles are published within the Clinton/Lewinsky lineage which in turn, on average, make more use of the Lewinsky trait. Thus, the Lewinsky noun phrase, as it appears in print, is the expression of a lexico-syntactic active replicator.

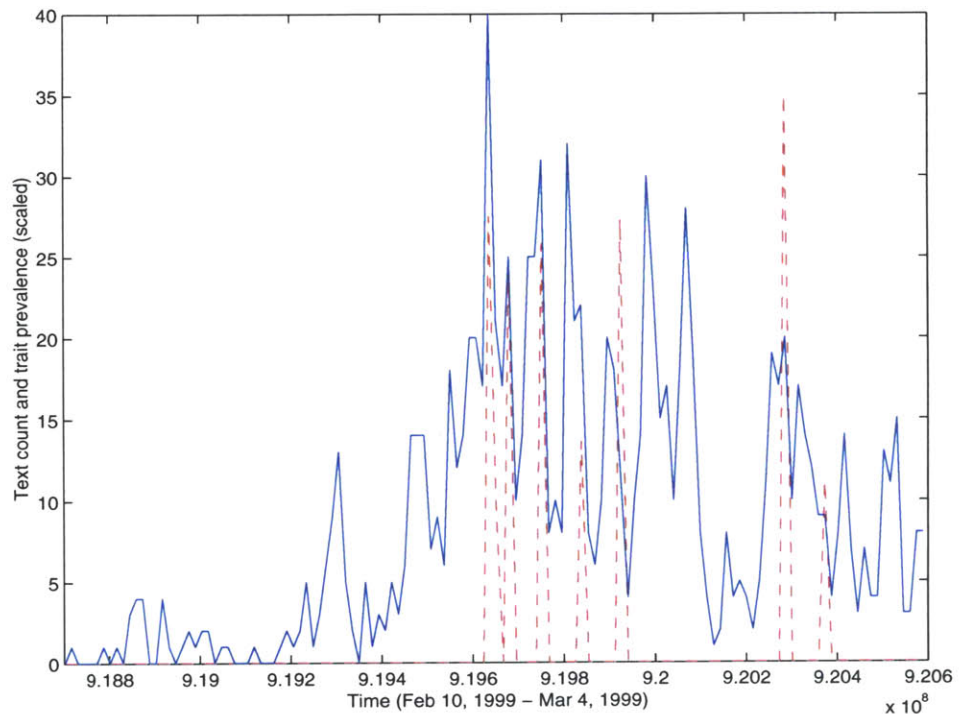


FIGURE 9. The “right wing ignorance” trait (scaled, dashed) and number of articles posted to the Clinton acquittal thread within the alt.politics.clinton NetNews newsgroup ($r = 0.42, p < 0.00001$). Texts are bucketed at four hour intervals; each point represents the number of posts within four hours.

I have done a similar analysis of the Clinton chronicon and have discovered some active lexico-syntactic replicators within these texts as well. Figure 9 describes just one of these active replicators. The noun phrase “right wing ignorance” occurs 239 times throughout the Clinton chronicon. The solid line shows the total number of posts to a large cluster of texts from the collection. This cluster comprises a very heated debate, again on the Clinton/Lewinsky scandal, dealing primarily with the

acquittal of Clinton in the Senate on all Articles of Impeachment brought against him. This acquittal occurred on February 12, 1999, which places this event prominently within the time period of the chronicon. Again, we see a clear positive correlation between these two timeseries ($r = 0.42$, $p < 0.00001$). And therefore, *ceteris paribus*, the more a text uses the noun phrase “right wing ignorance,” the more subsequent posts will appear within the same lineage. Thus “right wing ignorance” expresses an active replicator.

In Table 1, I offer more examples of active replicators at three different linguistic levels from the same two chronica, which gives a sense of the range of language phenomena that enjoys autocatalysis. These and other active language replicators will form the core of the results in support of the Microevolutionary Language Theory.

Chronicon	Cluster topic	Example replicators
Lexical replicators		
Globe	Clinton/Lewinsky	moral, public, denial, swear, true
Globe	Conflict with Iraq	strike, force, diplomatic, Hussein, Clinton
Clinton	Juanita Broaderick	lie, cunt, guilty
Lexico-syntactic (noun phrase and SVO) replicators		
Globe	Clinton/Lewinsky	Lewinsky sexual relationship, his personal life
Globe	Conflict with Iraq	US military strike, Iraq military action possible
Clinton	Gun control	concerns about Federal interference, inhibition of Second Amendment
Clinton	various	Clintonphobe grind tooth, it fuck wad, Congress make law
Syntactic replicators		
Globe	Clinton/Lewinsky	V, PRON, PRON, V (autocatalysis in question)

TABLE 1. Assorted active replicators from the Globe and Clinton chronica.

2.2 Summary

To recap, texts are clustered into topically related groups or lineages. Linguistic features (e.g., noun phrases) are distilled from the texts. Some features are found which replicate with significant frequency (e.g., hundreds of times) across the text collection, of which some positively correlate with the relative abundance of texts within their lineage. These, I argue, are active replicators, as their correlation with lineage population demonstrates autocatalysis.

These examples may have raised more questions than they have answered: How exactly do I have a lineage? How do I account for statistical artifact? What does active replication really have to do with the evolutionary process? I will attempt answers to all of these questions, and more, in the subsequent chapters.

The Chronica and CAMEL Software System

In this chapter I will describe in some detail three collections of timestamped texts that I call *chronica*. Totalling more than 17.5 million words, these text collections represent a significant data source (Sinclair, 1991).

I will detail the overall software system responsible for classifying the texts, revealing replicators, discovering active replicators, and so forth. Named CAMEL, for ComputAtional MicroEvolutionary Language, this set of programs forms the core system for all of my experiments. At nearly 20,000 lines of custom code, CAMEL represents a substantial piece of programming. Those portions of the software system which are specialized to particular experiments are detailed in Chapter 4 and elsewhere, but the general system is described here.

I conclude this chapter by applying the CAMEL system, as well as a collection of specialized algorithms, to compute popular stylostatistical features of the *chronica*. This will give us a general sense of these texts, as well as suggest future directions for comparative corpuslinguistic studies of internet discourse.

3.1 What Is a Chronica?

This dissertation explores the course of language over time and the accumulation of design at simple levels. Such an exploration, at least at the scale conducted here (small in time, yet broad in data coverage), would not have been possible even ten

years ago, due in part to the heavy computational requirements of the system. Happily, sufficiently powerful desktop workstations have come to the rescue. But this work is also made possible thanks to the recent and significant rise in online collections of text. UseNet News (NetNews) discussions, newspaper and magazine archives, e-mail discussions, listserves, manuscript traditions, scientific papers, web pages, and so forth, all provide ideal corpora for the type of analysis in which I am engaged. All of these corpora have a clear arrow of time running through them: The individual texts are consistently timestamped; the authors are, in general, sensitive to the temporal context (e.g., newspapers are “timely,” NetNews authors know what recent posts have been made). Such collections I refer to as *chronica* (*chronicon* in the singular) which I take from the Greek root for words such as “chronicle.”

These text collections are different in many ways from more traditional corpora used in corpuslinguistic analysis, and I’ve coined this neologism to help emphasize these differences. To understand this distinction it is useful to consider two of the most famous text collections within the corpuslinguistic community: (1) The Brown corpus exemplifies a standard collection of texts (Kucera & Francis, 1967; Francis & Kucera, 1979; Kucera, 1992). It is composed of approximately one million words of text in American English, published in 1961. The goal of the Brown corpus is to synchronically capture American English usage for that year. Though the corpus does indeed contain time-sensitive prose, such as newspaper articles, the dynamic of time through the collection itself is consciously discounted. (2) The Lancaster-Oslo/Bergen Corpus (LOB), another quite famous corpus, was assembled with the goal of serving as the British counterpart to the Brown corpus. It also consists of one million words of text, this time of standard written British English, that was published in 1961 (Johansson, Leech & Goodluck, 1978). Both the Brown and the LOB corpora represent traditional collections of texts within the corpuslinguistic community; I would not refer to them as *chronica*.

3.2 Globe Chronicon

My experiments have relied on three primary *chronica*. The first, which I call the “Globe” collection, is composed of 22,498 articles printed in the National and International News sections of *The Boston Globe* in 1997 and 1998.¹ These texts required extensive pre-processing and reformatting. Each article includes a title, date of publication, author information, and some keywording performed by the *Globe* staff. (I made only minimal, if any, use of this keywording in my analysis.)

1. Lisa Tuite, Librarian of *The Boston Globe*, and her staff receive my sincere thanks for providing this corpus.

Figure 10 is an example of an article from the chronicon after being pre-processed. Notice that the date is given both in a standard format (12/7/98) and in the Unix format (913006800), which is the number of seconds since the Epoch January 1, 1970.

```
TITLE: LIBYANS OFFER HINTS OF A SOLUTION TO LOCKERBIE IMPASSE
DATE: 12/7/98 913006800
SOURCE: By John Daniszewski, Los Angeles Times
KEYWORDS: US LIBYA TERRORISM RELATION NAME-LOCKERBIE TRIAL
TRIPOLI, Libya -- Perched on a fence above the city's
seafront, with nothing better to do all day than watch the cars go
by, the jobless man showed no hesitation when asked whether
his government should surrender two suspects wanted in the
bombing 10 years ago of Pan Am Flight 103. "The Lockerbie
case should be resolved and those two men should be
extradited," said Khaled Sadq, 31, a university graduate who has
remained unmarried because he said he could not find work to
support a spouse....
```

FIGURE 10. Example of an article from the Globe collection. All articles include title, date (in standard and Unix formatting), sourcing information, keywords, and article body.

3.3 Overview of NetNews

The other two primary chronica come from posts to NetNews newsgroups. NetNews offers an excellent source of texts for microevolutionary analysis. It originated in 1979 as a software mechanism to distribute among computers connected to the early internet "bulletins, information, and data... items of interest such as software bug fixes, new product reviews, technical tips, and programming pointers, as well as rapid-fire discussions of matters of concern to the working computer professional," (Kantor & Lapsley, 1986, Section 1). This distribution was for the benefit of the ARPA-internet community and within the first year fifty UNIX sites were participating. Like all of the internet, NetNews is defined solely by its protocols (rather than by ownership or licensing or governance). The Network News Transport Protocol (NNTP) stipulates how NetNews messages are posted, distributed, and retrieved over the internet (Kantor & Lapsley, 1986). A further internet memo specifies the actual format of each NetNews message (Horton & Adams, 1987).

The collection of messages over NetNews is organized into subject groups, called *newsgroups*, which, in turn, are organized in a tree-like hierarchy. At the top of the hierarchy is a collection of broad categories, most notably:

alt	Alternate groups
bit	Gatewayed BITNET mailing lists
comp	Computer professionals and hobbyists
misc	Groups not fitting anywhere else
news	USENET News network and software
rec	Hobbies and recreational activities
sci	Research/applications in established sciences
soc	Social issues and socializing

Top-level categories of local interest may also be created: for instance, “mit” for news items of interest to the MIT university community. Underneath each top-level category are newsgroups as well as possible further hierarchical categorization. A newsgroup name is defined as the entire path from the top-level category through any subsequent refining categories down to the name of the group itself. Category and group names are delimited by the period symbol. Thus, “sci.physics” is the name of a scientifically oriented newsgroup devoted to general physics subjects. However, “sci.physics.plasma” is a more specific group devoted to the study of plasmas (see Figure 11). A voting mechanism exists in which new groups are proposed and approved for addition to NetNews (though groups within the “alt” domain require no vote to be created). Today there are thousands of newsgroups dealing with every possible subject matter.

Users access NetNews through one of any number of news reading software systems. The systems all offer a few essential features: A user subscribes to those newsgroups that are of interest to them; the news reader keeps track of messages sent to these newsgroups and notifies the user when new messages have arrived; the user can read posts sent to these newsgroups and can post new messages as well.

Posted messages are transmitted, via the NNTP protocols, to other USENET users across the Internet.

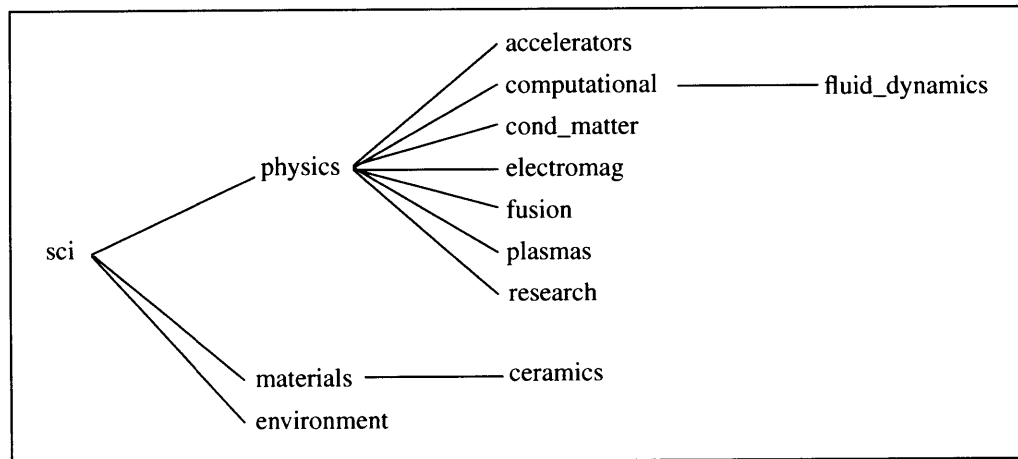


FIGURE 11. A small portion of the NetNews newsgroup hierarchy.

Posts are composed of a number of fields, of which only a few are relevant here. The user creating the post is responsible for the post body, that is, the actual text of the message, as well as a subject line. The subject line is composed of a few words which describe what the post is about. NetNews software will append to posted messages a number of additional fields including a timestamp and the user name of the person who created the post.

Posts can be either an independent message or a follow-up to a previous message. A follow-up, or “in-reply-to” message, will have special threading information in its header linking it to the previous posts to which it is a reply. This header information allows news readers to reconstruct the discussion thread. Further, in-reply-to messages will, by default, share the same subject line as the original message (though the poster of a follow-up can choose to change this).

Users who post to NetNews often send their message to a single newsgroup. However, it is possible to send a single message to multiple newsgroups. Called “cross-posting,” this is generally done when a message is relevant to multiple lists and is a way to broaden the potential readership.

3.4 Netnews Chronica

I have made use of two different chronica taken from NetNews postings. The “Clinton” collection is composed of 15,702 texts posted to the alt.politics.clinton newsgroup during the first few months of 1999.¹ Posts to this newsgroup deal with U.S. President Bill Clinton. The period of time captured by this chronicon is at the height of Clinton’s impeachment trial in the U.S. Senate. The charges brought against him centered on his attempts to cover up his sexual relationship with a former White House intern, Monica S. Lewinsky. Clinton was acquitted on charges of perjury and obstruction of justice by the Senate on February 12, 1999. In Figure 12, I show a sample post from the alt.politics.clinton newsgroup. The fields shown

```
Subject: Re: Do the Republicans and Independents Deny It?
Date: 918859369
From: Larry Smith <jlaw@bellsouth.net>
Newsgroups: talk.politics.misc,alt.rush-limbaugh,alt.fan.rush-limbaugh,
alt.politics.clinton,alt.current-events.clinton.whitewater

Anthony Stephen Szopa wrote:
>
> Do the Republicans and Independents Deny It?
>
> Do the Republicans and Independents deny that clinton has assaulted
> you, the Constitution of the United States, the Rule of Law,
> and trashed the White House?
>
> Do the Republicans and Independents deny that the Democrats and
> clinton defenders held you down as clinton politically raped you?
>
> Now that you are fucked, how do you ever expect to get unfucked?

Take a sedative and sleep it off.
```

FIGURE 12. Example post from alt.politics.clinton. This is a in-reply-to post with the previous message’s text delimited by the “>” symbol. Note that this text was cross-posted to five different newsgroups.

1. My thanks to Doug Bagley of DejaNews for his assistance in acquiring this collection.

are: subject line, date (shown in Unix format), author information, newsgroups posted to (this text went to five different newsgroups), and the post body. Note that this example is an in-reply-to post. Some of the original message is preserved, and is delimited in the body of the message by the ">" symbol. The new text consists of simply one sentence: "Take a sedative and sleep it off."

The second NetNews chronicon, which I call "Skeptic," consists of 11,758 posts primarily to the sci.skeptic newsgroup during September of 1995. Some other newsgroups within the sci.* hierarchy also are represented in this collection. The sci.skeptic newsgroup is composed of posts on scientific issues, with a skeptical attitude. In Figure 13, I show an example post from this chronicon. Note that it, too, is an in-reply-to message. Furthermore, note that the date field is shown in yet another format among the handful of standard date formats found in NetNews posts.

```
Subject: Re: ALIEN AUTOPSY/FOX
Date: 24 Sep 1995 16:23:37 -0400
From: lazzwaldo@aol.com (LazzWaldo)
Newsgroups: sci.skeptic
```

Alan Barclay sez:

```
<I already said back in this thread that I don't believe it was a
<real alien. I'm only debating the assumptions being made to
<debunk it.
```

```
What's wrong with the using the assumption "it's highly unlikely"
to help debunk it? was a real alien, WHY don't you believe it was
real? If you don't believe it, can we use those same assumptions
in our debunking, or are YOUR assumptions intellectual property
we can't use?
```

FIGURE 13. Example post to sci.skeptic newsgroup, also an in-reply-to message.

3.5 Summary of Chronica

These three chronica, taken in total, amount to nearly 50 thousand texts and over 17.5 million words. Among standard collections in the corpuslinguistic community,

this is large, though not the largest (some very recent corpora have contained over 100 million words) (Jane & Lampert, 1993). Nonetheless, this is a substantial dataset, adequate for studying salient patterns of usage and meanings (Sinclair, 1991).

Chronicon name	Number of texts	Number of words	Date range
Globe	22,498	7,906,642	3/1/97 - 12/16/98
Clinton	15,702	5,843,958	1/29/99 - 3/5/99
Skeptic	11,758	3,820,878	9/20/95 - 9/26/95
Total	49,958	17,571,478	

TABLE 2. Basic characteristics of three chronica. Globe chronicon is composed of articles published in *The Boston Globe*, Clinton chronicon is composed of posts to alt.politics.clinton, and Skeptic chronicon is composed of posts primarily to sci.skeptic newsgroup.

3.6 CAMEL System Overview

The base CAMEL software system relies on a collection of techniques that are mostly well known within the information retrieval and analysis communities (Frakes & Baeza-Yates (1992) provide a fairly current review). However, the goals are different from those of traditional IR systems: I wish to classify documents into lineages, distill replicators, and determine which replicators are active. The initial steps in this analysis are identical, regardless of the particular investigation, and are illustrated in Figure 14. The final set of steps varies depending on the linguistic level studied (e.g., lexical versus syntactic); these differences will be described in Chapter 4. The goals for these initial steps are: To perform a preliminary analysis of the documents; to compute for each text a numeric vector representation that posi-

tions it in some sort of conceptual space; and to classify the texts into related groups.

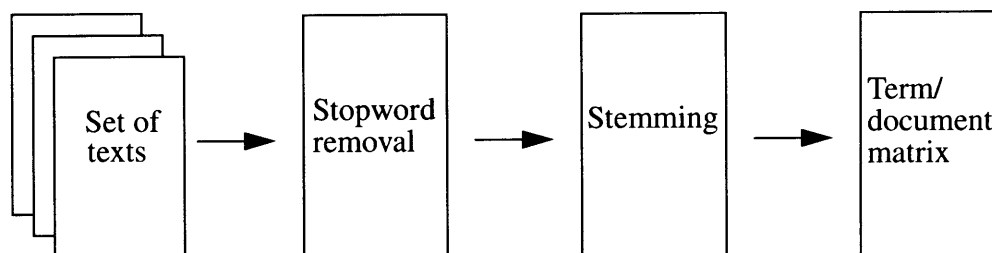


FIGURE 14. Initial steps in CAMEL text analysis. These first four steps are performed on all chronica regardless of the particular experiment.

3.7 Stopwords and Stemming

The first step in the preliminary analysis of the documents is to assemble each chronicon into a collection of files, one for each text. For each individual text the system will then analyze the words from both the body and the subject line or title; all of these words are combined and form a single word list. Those words that are so common in English as to carry little or no semantic content are removed from the list: for instance, function words, pronouns, and other common words such as “and,” “is,” “I.” This collection of common words is generally called a *stoplist* and individual words are referred to as *stopwords* (Fox, 1992). My stoplist was developed in two steps. First the lexicon of Karttunen’s (1983) large, morphological analyzer and part-of-speech tagger was used to identify the function words in English. Next, a list of high frequency words was assembled from a word frequency analysis of the Brown corpus (Francis & Kucera, 1982). The resulting stoplist contains 270 common words.

Those words from a text that make it past the stoplist are then passed through a stemmer. The process of stemming attempts to conflate morphologically similar words into single terms by removing suffixes and prefixes and normalizing tense (so that “eating” becomes “eat” or “traveler” becomes “travel”) (Frakes, 1992b). The goals here are twofold: The strong semantic link between the various forms of a root word is made explicit; moreover, the total number of words that will need to be considered in subsequent analysis is reduced. I currently make use of two different stemming systems. The first, due to Porter (1980), is a strictly rule-based algo-

rithm which locates and strips suffixes (such as “ing,” “ed,” “s”). The algorithm does not attempt to correct word spellings after the suffix is stripped (for instance “running” will become “runn”). However, this should not be a problem or overly reduce the benefits of stemming which demands only consistent application of stemming across all words (A.F. Smeaton, personal communication, 1995).

The other approach to stemming uses an English lexicon and morphological feature system named ENGTWOL (Heikkilä, 1995a; Heikkilä, 1995b). This mouthful is an abbreviation for “English two-level,” in reference to Koskenniemi’s (1983) model of the same name. The version I use is embedded in a commercial constraint-based morphological analyzer and part of speech tagger, EngCG-2 (Samuelsson & Voutilainen, 1997), which, as the name suggests, is a second-generation version of the original EngCG (Karlsson, Voutilainen, Heikkilä & Anttila, 1995). The primary goal of EngCG-2 is to tag words with their part of speech and inflectional properties. It accomplishes this by employing both a large lexical database with over 56,000 entries and a grammar of over 4,000 contextual rules (Conexor oy, 1998a) which is primarily designed to disambiguate those words assigned multiple grammatical tags. This process also assigns to each word a *lemma* derived from the ENGTWOL stem lexicon. The lemma is the word’s lexeme, or stem word, abstracted from any of its various word forms (e.g., “run” for “running”). However, if the word is not located in the lexical database, an heuristic stemming system is applied, similar to Porter, in order to arrive at a lemma (Voutilainen, 1995a). This overall scheme is far more accurate than the Porter system, though it is computationally costly and relies on a large grammar and lexicon.

When the stemming process is complete, all of the words from all the texts in the chronicon are combined and lexicographically sorted, removing any duplicates. Simultaneously, the system counts the number of times that each particular word occurs across the entire collection of texts. In a continuing effort to shorten this list of words, any words which do not occur in a minimum number of texts are removed; this minimum threshold is currently set at three. This step helps to remove uncommon misspellings, nonsense words, and words that are so obscure they have no discrimination value. The final set of sorted, stemmed, and pruned words makes up what is called the *term list*.

3.7.1 Example

Consider the NetNews text in Figure 15.

```
Subject: Re: wrecks to locate?
Date: 26 Sep 1995 04:08:22 -0700
From: joe@foo.com (James Smith)
Newsgroups: sci.military.naval

bill@osl.or.gov (Bill Smith) writes:
>Should all of the ships sunk at Pearl Harbor and
>Tranto been left there because they went down
>with casualties aboard?

Arizona HAS BEEN left at Pearl Harbor...
The others were salvageable and the US used them to fight the
Japanese.
James Smith
Cypress, CA
```

FIGURE 15. Text of a fictitious post to sci.military.naval. The emboldened text represents those parts used by the CAMEL system (note that quoted material is counted in analysis)

This message is in-reply-to an original posting by Bill Smith and the first three lines of the text body are quotes from Smith's original post (the quoted text can be identified by the ">" symbols). Only the subject line and the text of the posting are passed through the stoplist and stemmer. The resultant list of terms, using the Porter stemmer, often results in nonsense words. The term list for this text is:

aboard	locat
arizona	other
bill	pearl
ca	re
casualti	salvag
cypress	ship
fight	smith
harbor	sunk
jam	went
japanes	wreck
left	write

If ENGTWOL is used to lemmatize the text instead, the word list is different. ENGTWOL not only will normalize inflected forms, but also will lemmatize irregular verb forms and normalize number and tense. Using the emboldened words from Figure 15 as input, the string of lemmatized terms, without punctuation, output by the ENGTWOL system is:

```
Re wreck to locate should all of the ship
sink at pearl harbor and tranto be leave
there because they go down with casualty
aboard Arizona have be leave at pearl har-
bor the other be salvageable and the US
use they to fight the Japanese James smith
cypress ca.
```

And the resultant term list, removing the stopwords, is:

Arizona	harbor
James	leave
Japanese	locate
Re	other
US	pearl
aboard	salvageable
because	ship
ca	sink
casualty	smith
cypress	tranto
fight	wreck
harbor	

3.8 Vector Space Representation

The term list is the key to creating a vector representation for each post. Consider a list of n terms distilled from a corpus of texts. Each post is represented within this n -dimensional vector space as a length n *term vector*. This vector space representation has been used extensively within the text retrieval community (Salton & McGill, 1983; Harman, 1992). It is necessary to select a particular method for computing the values for each vector; that is, how should we define the *term weighting*? The text retrieval community has shown that both the frequency of the term *within* a particular document, as well as the frequency of the term *across all* the documents,

is useful in determining the weight assigned to each element in the vector (Salton & Buckley, 1988).

The within document frequency, or *term frequency*, acts as a good measure of the salience and relevance of a term within a text. Simply put, the more frequent a word occurs in a document, the more relevant the word is to that document. As is generally the case, the term frequency is determined by summing the total number of occurrences of a term in a text. However, in the CAMEL system, those words from the subject or title are given a higher weighting than those from the text body, reflecting their higher discrimination value.

It is also common practice to include information about the overall frequency of the terms across the entire text collection in the weighting. If a term is extremely frequent across a chronicon it is unlikely to have high discrimination value even though it may also be of high frequency in a particular text. For computing the collection-wide term frequencies we use the well-known *inverse document frequency* (*IDF*) (Salton & Buckley, 1988). Consider a chronicon of m texts and a particular term, j , within the list of n terms. Then the IDF is given by

$$IDF_j = \log\left(\left\lfloor \frac{m - m_j}{m_j} \right\rfloor\right).$$

Here m_j is the number of posts in which term j appears and $\lfloor \cdot \rfloor$ represents the integer floor operator. The term weight for a document, i , and term j is defined by

$$TermWeight_{ij} = w_{ij} = \log(TermFrequency_{ij}) \cdot IDF_j.$$

Each term weight, then, is a combination of its inter- and intra-document frequencies. This particular term weighting scheme is commonly referred to as TF/IDF.

Each text, i , is now represented by a particular term vector,

$$r_i = (w_{i1}, w_{i2}, \dots, w_{in}).$$

The entire collection of m term vectors, one for each text, defines the *term/docu-*

ment matrix, A ,

$$A = \begin{bmatrix} r_1 \\ r_2 \\ \dots \\ r_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix}.$$

3.8.1 Example

I offer a small example to illustrate how the term/document matrix is computed. Consider a corpus of three texts with the following very small term lists. The number next to each term is the term frequency; the number of times the term occurs in the text.

text 1	text 2	text 3
harbor 2	harbor 2	left 2
japan 1	pearl 2	locat 1
pearl 3	salvag 4	write 4
ship 2	sunk 2	

It should be easy to verify that the term list for the entire corpus has nine entries: harbor, japan, left, locat, pearl, salvag, ship, sunk, write. Again, thanks to using the Porter stemmer, some of these terms are not English words. (Note: for this example, a term is not required to occur in more than two texts to be included in the term list.) The equation above describes how to compute the IDF for a term. The resultant IDF for each term (where $\log(0) \equiv 0$) is:

harbor 0	salvag 1
japan 1	ship 1
left 1	sunk 1
locat 1	write 1
pearl 0	

For each text the term weights are computed by multiplying the above IDF values by the integer floor of the base 2 logarithm of the term frequencies. The resultant

term/document matrix is given by

$$A = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}.$$

Here the rows represent the nine terms alphabetically from left to right. Admittedly, the size of this example produces a term/document matrix with little clear discrimination power.

3.9 Text Clustering

The next step, given a term/document matrix, is to classify all of the documents according to their content (Figure 16). This is accomplished by clustering the vector space representations. This is a text classification method similar to that used by text analysis systems, such as web search engines (Rasmussen, 1992). I am currently employing the nearest neighbor algorithm to perform this clustering (Jain & Dubes, 1988). This method is natural, simply implemented, and computationally tractable. The algorithm requires the initial input of a distance threshold, t . This threshold specifies the maximum allowable distance between two vectors assigned to the same cluster.

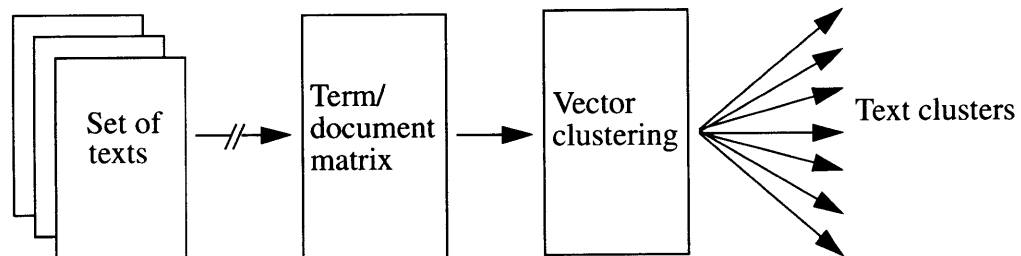


FIGURE 16. Clustering step.

The algorithm considers in turn each row vector from the matrix, A . The current vector is compared to each vector already assigned to a cluster. If the closest of such vectors is not farther than the threshold distance, then the current vector is assigned to that cluster. Otherwise the current vector is assigned to a new cluster. For the experimental chronicon I have discovered an appropriate threshold, t , through trial and error. I use this same distance threshold for all of the principal

chronicon. In the future I might employ techniques which do not require a user supplied threshold.

Determining whether two vectors are “close” to one another is not a trivial matter. A collection of metrics have been proposed for use with text clustering (Jones & Furnas, 1987; Salton & Buckley, 1988). I am currently using a similarity measure well known within the text-retrieval community, namely, the *cosine* measure (Figure 17).

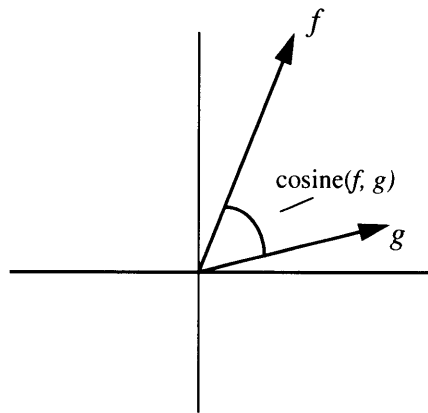


FIGURE 17. The cosine measure.

The measure is aptly named since it computes the cosine of the angle between its two arguments. Consider two documents represented by their n -length vectors, f and g . The cosine measure is defined as

$$\text{cosine}(f, g) = \frac{\sum_{i=1}^n (f_i \cdot g_i)}{\sqrt{\sum_{i=1}^n (f_i)^2 \cdot \sum_{i=1}^n (g_i)^2}}.$$

The numerator should be recognized as the vector inner product. The denominator is the product of the vector norms and thus serves as a length normalization term. It should be noted that the row vectors of A are not normal. In fact, their vector norm is a function of the length of the post. I do not want the similarity measure to be

sensitive to text length, therefore, the additional normalization term of the cosine measure is important. (Elsewhere (Best, 1996) I've criticized cosine for not acting strictly as a metric: It is not mathematically admissible as it violates the triangle inequality.)

3.10 Correlation Coefficients

So far the CAMEL system has processed each text, determined the term frequencies, coded a vector representation for each document based on the term frequency and inverse document frequency, and clustered the texts into conceptual groups based on these vector representations. Note that this is already enough to generate the graph of Figure 7 in Chapter 2 by simply counting the number of texts within a cluster week by week and graphing against time these weekly sums (since all of my texts are timestamped).

In Chapter 1, I introduced the notion of replicators and active replicators and argued that in evolving systems active replicators deserve our highest attention. The CAMEL system has already distilled and tracked a potentially useful set of replicators from the chronicon: lexical replicators. The term frequency analysis described in Section 3.8 computes the rate of reoccurrence for terms (word lemmas) across the time-valued collection of texts. These are replicating traits, much the same as the lexico-syntactic replicators offered as examples in Chapter 2. But, obviously, they are much simpler to distill.

Given some replicating trait over time, how can I determine if the trait is active? My method is identical regardless of what the actual replicator is (word, noun phrase, etc.). So the technique described here for lexical replicators will be identical to the technique applied for other replicators. My goal is to find examples of replicators that are autocatalytic; in other words, they must have high correlation with the publication (or post) volume within some cluster. This is the same sort of linking of a trait to success seen in Chapter 1. In Chapter 2, I used the Pearson product-moment correlation coefficient (r) to measure quantitatively how closely two timeseries were correlated. The Pearson correlation is a standard measure of relatedness between two variables and is reviewed in any standard statistical text (e.g., Howell, 1995). The value for r varies between -1.00 and +1.00. Values near 0.00 attest to variables that have little or no relationship. Values approaching -1.00 are negatively correlated; rises in one variable tend to occur with drops in the other variable. Similarly, values for r that approach +1.00 describe variables positively correlated; here, rises in one variable tend to occur along with rises in the other variable.

For some high value of r , the relative presence of the linguistic feature is correlated strongly with the volume of published texts within the given cluster. A high enough r shows that this replicating feature is active; its presence is catalyzing the appearance of texts within the topical cluster and these texts contain with high probability the same feature. This is autocatalysis and replicator power (Blackmore, 1999)!

But how big an r is enough? This is currently left as a subjective decision. For each level of replicator, I have chosen an r that I feel is sufficient to label the replicator active. But this binary decision is a bit unnatural; clearly, some replicators will have higher correlations than others and this describes a continuum of autocatalytic affect.

3.11 Accounting for Statistical Artifact

To measure and interpret correlation coefficients fairly is a tricky business indeed. Even two completely random variables might easily correlate. To avoid statistical artifact any potential for spurious correlations must be accounted for. While developing the correlation analysis system within CAMEL, I found two areas of potentially artificial correlations between the volume of published texts within a cluster and some linguistic replicator. The first is zero-points. Clearly, if there are no texts published to the cluster during some time period, there will be no replicating traits within those texts, guaranteeing that these points will correlate, as they both will be zero. I account for this correlating artifact by removing all zero-points from both timeseries; thus I compress out the points in time when no texts are published.

The second area of artificial correlation which I empirically discovered was due to text volume. It turns out (somewhat to my surprise) that the average size of a set of texts correlates lightly with the volume of those texts. In other words, as the number of texts published to some cluster goes up, the length of the texts also has a tendency to rise. This is easily accounted for by simply normalizing the traits by the size of each text. When computing correlation coefficients, the traits represent the relative proportion of text devoted to the feature under consideration. These two precautions seemed to produce series free of artificial correlations.

Now, given two series that we believe to be free of spurious correlations the remaining question is: How much is enough? In other words, how correlated do two variables need to be in order to argue that the relationship is significant? One way to approach this problem is to argue against the standard null hypothesis: the two variables are not significantly correlated. Given some correlation coefficient, and a sample size of n , we use a standard two-tailed test to determine the probability that

the null hypothesis is to be accepted. These probabilities, p , are reported (as was done in Chapter 2) along with the correlation coefficient. Thus, the expression $p < 0.00001$ says that we should accept the null hypothesis with a probability less than 1 in 100,000.

But even this may not be good enough, in particular when one is doing a lot of statistics. For instance, if I'm measuring 100,000 correlation coefficients then I might expect something that happens 1 in 100,000 times to indeed occur. In other words, given a large number of correlation coefficients we need to make sure that we are not simply sampling the tail of some distribution about 0.00. For the results reported in the next chapter I've computed the mean and standard deviation about the mean for the set of correlation coefficients I discuss. In this way, I can see just

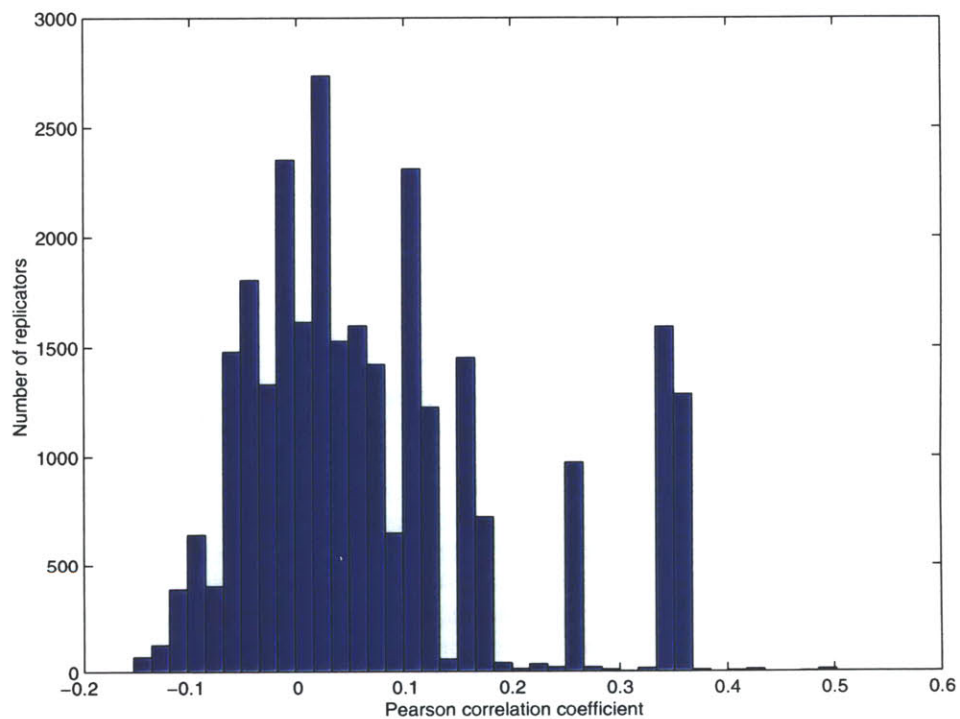


FIGURE 18. Histogram of correlation coefficients for all lexico-syntactic replicators from Clinton/Lewinsky articles within Globe chronicon.

how far out on the tail of a distribution the reported correlations lie. For any results I report on, the reported correlations are quite far out indeed. For instance, in Chapter 2, I offered a preliminary example, the “Monica S. Lewinsky” lexico-syntactic replicator, from the Globe chronicon. The mean correlation coefficient for lexico-syntactic replicators within this collection is 0.0736 and the standard deviation is 0.1258. That means that the correlation coefficient for the “Monica S. Lewinsky” replicator, $r = 0.44$, is indeed in the tail. In Figure 18, I have histogrammed the correlation coefficient values for all lexico-syntactic replicators within that cluster.

I have attempted to account for all potential sources of artificial correlation between these timeseries. Further, I’ve shown that while these strong correlations may be due to chance this is most unlikely. In the end, however, it is a *qualitative* assessment of the validity of these statistics that will be the most important. The proof of statistical efficacy will ultimately be predicated on whether I can weave convincing qualitative explanations as to why some replicators appear active whereas most do not. In other words, is there any conceptual explanation as to why some replicators are highly correlated (and thus, I will argue, of evolutionary significance) and some are not?

To strengthen my case, and to offer a specific example, consider the lexical replicators distilled from the Globe chronicon. In Table 3, I show the five most highly correlated active lexical replicators for the Clinton/Lewinsky scandal cluster of texts (see Section 3.12) along with the five most uncorrelated replicators. I’ve listed the correlation coefficient, r , along with its value for p , and the number of times the replicator appears within the chronicon, n . The first five rows in Table 3 describe replicators which have high correlation with the level of Clinton/Lewinsky reporting; the last five rows are the most uncorrelated.

My question is this: does “moral” have more significance to this population of Clinton/Lewinsky stories than “read”? Morality is one of the central issues and hot-buttons in this story. The lemma “read” appears more often than “moral,” but conceptually is of small importance. I believe that Table 3 provides strong qualitative support that these active replicators are not simply the result of statistical artifact but instead represent true semantic socio-cultural dynamics. The lexical replicators with high correlation are those words that might be expected to have evolutionary significance; whereas, this is not true for the uncorrelated replicators.

<i>r</i>	<i>p</i>	<i>n</i>	Lexeme
0.6333	< 0.000001	51	moral
0.5866	< 0.000001	408	public
0.5537	< 0.000001	60	denial
0.5391	< 0.000001	768	president
0.5334	< 0.000001	73	swear
0.0002	≈ 1	70	read
0.0001	≈ 1	59	reveal
-0.0002	≈ 1	143	head
-0.0003	≈ 1	79	Indiana
-0.0004	≈ 1	72	1993

TABLE 3. Five most correlated lexical replicators and five least correlated from Clinton/Lewinsky articles within Globe chronicon.

The mean correlation coefficient for all lexical replicators within the set of Clinton/Lewinsky articles is $r = 0.0840$ and the standard deviation is $\sigma = 0.8978$. In Figure 19, I show the histogram for the correlation coefficient of all lexical replicators

within this group of texts. The large values shown in Table 3 are certainly far on the tail.

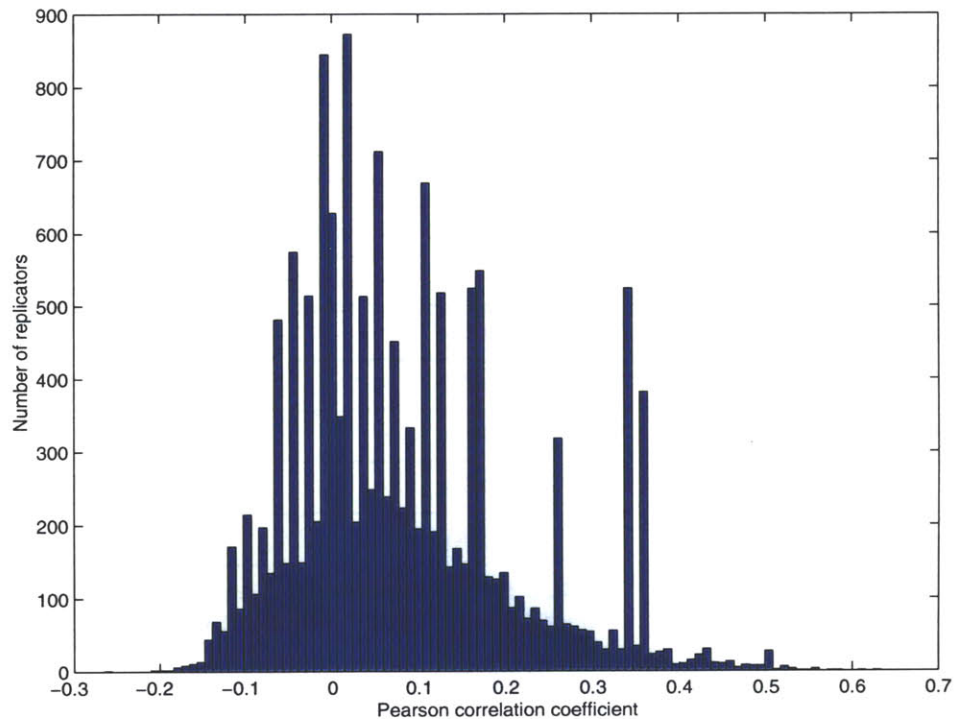


FIGURE 19. Histogram of correlation coefficient for all lexical replicators from Clinton/Lewinsky reporting within Globe chronicon.

3.11.1 Neutral shadow

Finally, I'd like to consider one additional test against statistical artifact. Table 4 shows the results of a neutral shadow model (Beadu & Packard, 1992; Beadu, Snyder, Brown & Packard, 1997) of the lexical replicators within the Globe chronicon.

To produce a neutral shadow for these lexical replicators I randomized each text's vector within the term/document matrix. That is to say, I took the original set of term vectors as described in Section 3.8 and randomly permuted the weights across the set of terms. In this way, I preserved the length of each vector and the Zipf dis-

tribution of words (see Baayen, 1993) while creating a collection of documents composed of randomly selected words.

This randomization process results in a collection of pseudo-texts composed of words selected at random (though drawn according to the original word distributions). Therefore, the distribution of correlation coefficients across these lexical replicators truly is the outcome of statistical artifact or some other structural properties of the texts that are not intrinsic to the replicators themselves.

I computed the Pearson correlation coefficient of all lexical replicators with text volumes from the original set of clusters. Table 4 shows the five lexical replicators with the strongest correlations and the five with the weakest for the Clinton/Lewinsky cluster. This should be compared against the same values originally computed for this cluster and displayed in Table 3. Clearly, the maximum correlation coefficient for the neutral shadow model is much smaller than those in Table 3. And the question here is: Does “moral” have more significance to this population of stories dealing with the Clinton/Lewinsky events than “month,” the most correlated feature from the random neutral model?

<i>r</i>	<i>p</i>	<i>n</i>	Lexeme
0.1997	< 0.0001	321	month
0.1270	< 0.02	405	office
0.1065	< 0.04	153	special
0.1064	< 0.04	49	cast
0.1044	< 0.05	47	sound
0.0004	≈ 1	149	change
0.0001	≈ 1	101	company
≈ 0	≈ 1	186	involve
-0.0001	≈ 1	113	Janet
-0.0002	≈ 1	167	send

TABLE 4. Five most correlated lexical replicators and five least correlated replicators for neutral shadow model of Globe chronicon, Clinton/Lewinsky cluster.

The histogram of the correlation coefficients for all lexical replicators within the Clinton/Lewinsky cluster under the random neutral shadow model is plotted in Figure 20. The average correlation coefficient is $r = -0.0409$ ($\sigma = 0.0158$). Clearly under the random model most all lexical replicators, distributed according to the same random variable, have an identical and slightly negative correlation with the volume of texts. I do not have a ready explanation for this negative value. But in any case it clearly supports the claim that strong positive correlations are not to be expected by chance nor to be due to systematic biases of the word frequency distribution, text length, temporal distribution, or the like (all things held constant under this random permutation).

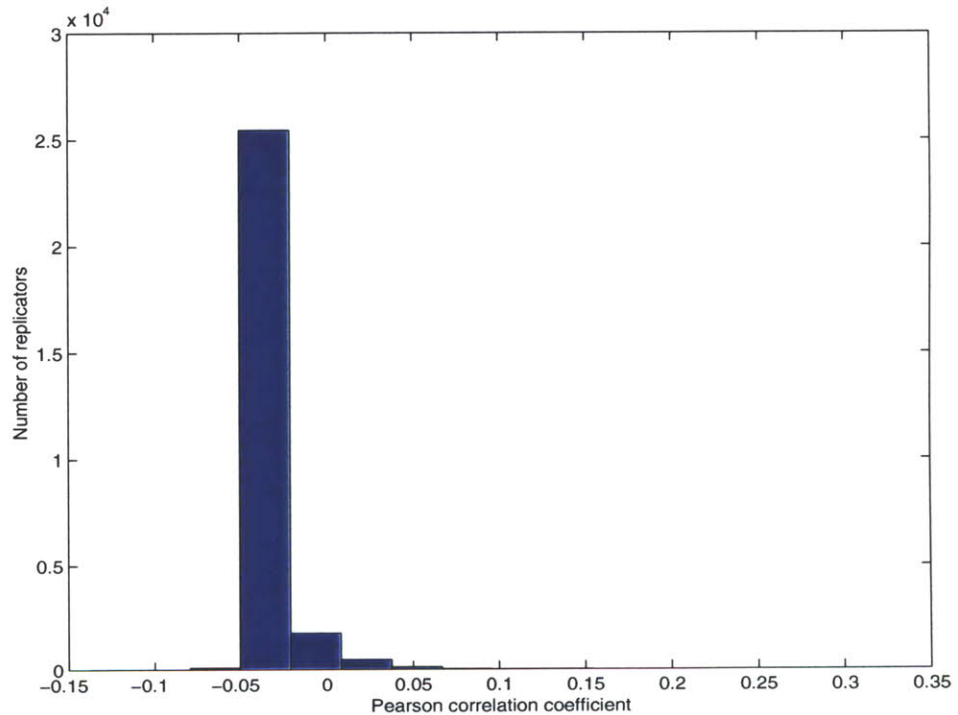


FIGURE 20. Histogram of correlation coefficient for all lexical replicators from neutral shadow model of Clinton/Lewinsky reporting within Globe chronicon.

There remains one other potential source of artifact that needs to be considered. Since the clustering process itself is a function of term frequencies, and these clus-

ters are used in turn to produce the timeseries that are then correlated, this could in itself produce correlations that are artifactual (R. Belew, personal communication, December 6, 1999). In other words, documents come together into some clusters because they share words in common. And that might be reason enough to produce correlations with the volume of texts within that cluster.

To test this theory I have produced a neutral shadow model with reclustering. Based on the Globe collection I created a new set of texts where each word was selected at random but according to the original Globe word frequency distribution. This random text collection also preserves the text lengths and temporal patterns from the Globe chronicon. I then reclustered all of these random texts using the same clustering method of Section 3.9. The reclustering step resulted in 1507 clusters with a mean cluster size of 15 texts ($\sigma = 44.4$) and a maximum cluster of 447 texts. Compare this to the original results shown in Table 8.

Table 5 shows the summary statistics for the correlation coefficients between text volume and lexical replicators for the three largest clusters. Figure 21 shows the histogram of correlation coefficients for the lexical replicators of the single largest cluster. And Table 6 shows the top five lexical replicators for the two largest clusters.

Cluster size	Mean r	Standard deviation of r	Max r
447	-0.0096	0.0560	0.2626
428	-0.0125	0.0572	0.2876
395	-0.0130	0.0546	0.2759

TABLE 5. Summarizing statistics for correlation coefficient, r , from three largest random clusters. Word frequencies and text lengths are same as Globe collection.

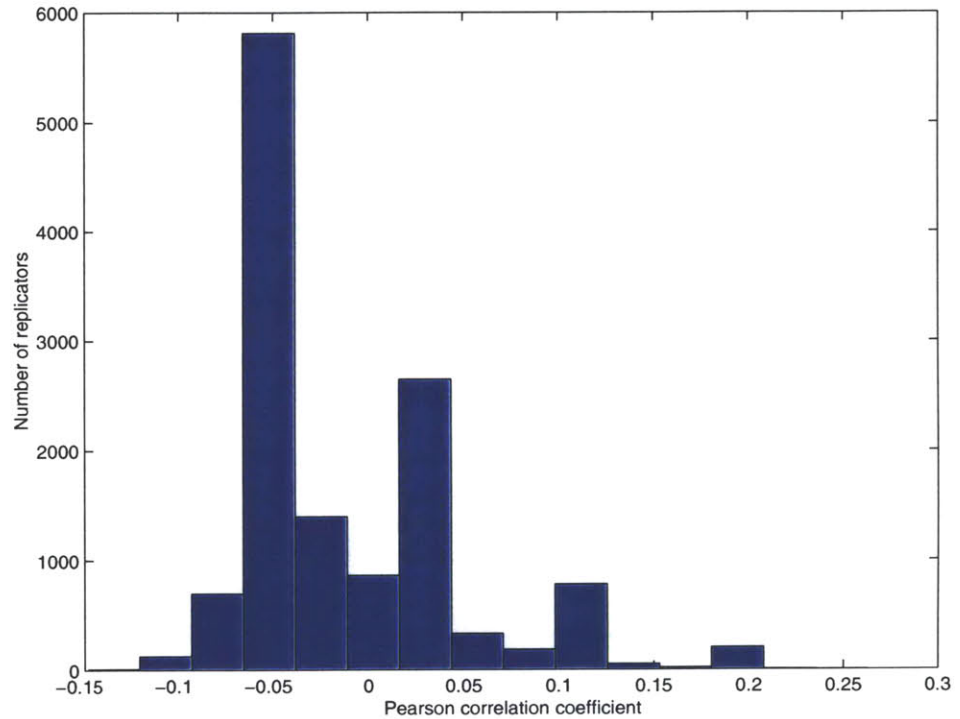


FIGURE 21. Histogram of correlation coefficients for all lexical replicators from neutral shadow model with reclustering. Image shows results from largest cluster.

What is clear from this data is that, as with the original neutral shadow experiment, this random model does not demonstrate any active lexical replication. It is interesting, though, to compare Figure 21 with Figure 20. Clearly, the clustering process does account for some small correlations (around $r = 0.2$), though in subsequent analysis I do not consider this low level of correlation to be of evolutionary signifi-

cance. I can conclude from this that any strong correlations ($r > 0.4$) between text volume and lexical replication is not due to an artifact of the clustering process.

Cluster Size	r	p	n	Lexeme
447	0.2625	< 0.0001	3	quash
	0.2563	< 0.0001	5	genetic
	0.2453	< 0.0001	2	castigate
	0.2119	< 0.003	2	Geraldine
	0.2119	< 0.003	11	bowl
428	0.2876	< 0.0001	2	consist
	0.2366	< 0.0001	2	retailer
	0.2328	< 0.0001	2	100-pound
	0.2302	< 0.0001	3	unpaid
	0.2248	< 0.002	4	rover

TABLE 6. Five most correlated lexical replicators from two largest clusters in re-clustered neutral shadow model.

3.11.2 Problems with timeseries

Timeseries are burdened with a few extra potential pitfalls of statistical artifact. They are notorious for evoking false positives from the Pearson coefficient (see McCleary & Hay, 1980; Gottman, 1981; or Wei, 1990 for overviews of timeseries analyses). A lot of statistics, including Pearson's, require that the timeseries be stationary. There should not be significant drift nor trend to the data. Trend is deterministic and drift is stochastic but both refer essentially to movement of the mean in some direction over time. To a large degree, this can be ascertained by simple inspection of the data (Gottman, 1981). Visually, the graphs of Chapter 2 and 4 do not suggest significant trend nor drift.

A more convincing demonstration that a timeseries is stationary comes from an inspection of the graph's correlogram (McCleary & Hay, 1980; Gottman, 1981). The correlogram is a plot of the k th-order autocorrelation against k for k ranging

from 0 across the length of the series, N . Here, the k th-order autocorrelation, rk , is computed for the timeseries f as

$$rk = \frac{\sum_{t=1}^{N-k} (f_t - \bar{f})(f_{t+k} - \bar{f})}{\sum_{t=1}^N (f_t - \bar{f})^2}.$$

A plot of this correlogram for a stationary series should not reveal any significant spikes, save that for $k = 0$.

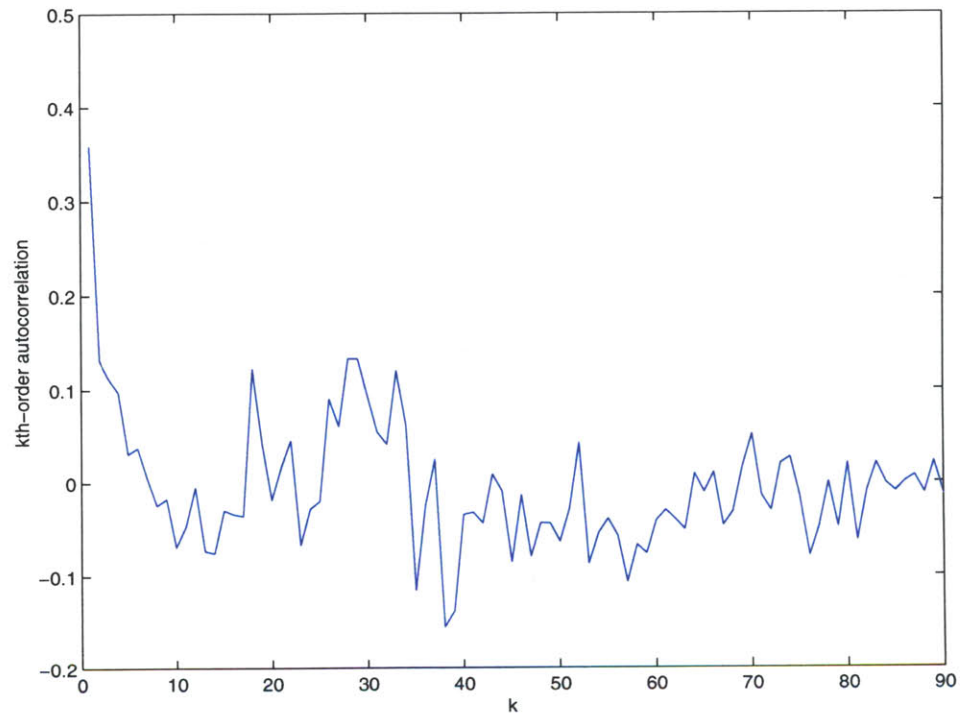


FIGURE 22. Correlogram for timeseries from the Globe chronicon. Timeseries represents number of articles published on Clinton/Lewinsky scandal in *The Boston Globe* (see Figure 7).

Figure 22 shows the correlogram for the Globe chronicon timeseries presented in Figure 10 of Chapter 2. This graph does not appear to be too spiky, save the expected high value for $k = 0$; things have a tendency to correlate well with themselves. A similar analysis of the other data in this dissertation shows reasonably stationary series as well

The other major bugbear when computing certain statistics, such as Pearson's, over a timeseries is due to autocorrelation in the series (McCleary & Hay, 1980; Wei, 1990). A timeseries is autocorrelated if it contains periodicities over time. A series that is not autocorrelated is said to be "white." The correlogram again is useful in determining this. But the best way to see if a dataset is white is to inspect its spectral domain for peaks. In other words, compute a discrete Fourier transform (e.g., an FFT) on the data. Again, you expect a peak at the 0th position due to the perfect autocorrelations at 0 time lag. But otherwise the graph should be symmetric about the middle and should not have any significant spikes. This would indicate data that is generally white and not autocorrelated. Figure 23 shows the spectral analysis for the same timeseries as in Figure 22. In the graph, I plot the frequency against the magnitude (real-valued portion) of the FFT. As hoped, it looks like white noise. This seems to be true for all of the timeseries I've studied; the conclusion is that no pre-whitening nor filtering of the datasets is required to account for autocorrelations.

Note that I have carefully chosen the bucket-sizes for each chronicon in order to ensure a pre-whitened timeseries. For instance, the bucket size of the Globe chronicon is seven days; this accounts for any inter-week autocorrelations which might be

expected for a newspaper (Monday's issue correlating with other Mondays, and so forth). The bucket size for the two NetNews chronica is four hours.

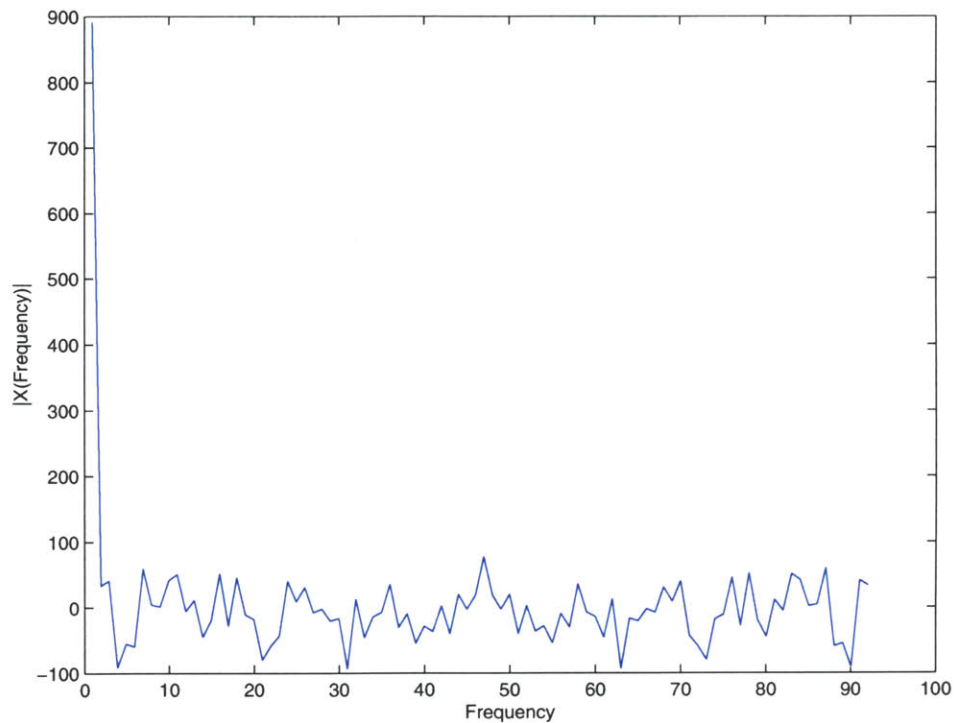


FIGURE 23. Fast Fourier Transform (FFT) of the timeseries of Figure 7. The spectral components are free from spikes and symmetric about the middle.

3.12 Chronica Revisited

The CAMEL system described above offers a number of insights into the nature of text chronica. For the Globe and Clinton chronica, I lemmatized the texts using the ENGTWOL lexicon. For the Skeptic chronicon, I used the Porter rule-based stemmer.

The resultant number of terms that make it through each of the steps described previously is shown in Table 7. This, along with the number of documents shown in

Chronicon name	Terms (w/out stop words)	Final termlist size	Ratio rejected	terms × docs (millions)
Globe	94,181	44,612	2.11	1,004
Clinton	41,500	28,185	1.47	442
Skeptic	41,980	27,031	1.55	318

TABLE 7. Size of termlist before and after removing of infrequent words, and final size of term/document matrix.

Table 2, describes the total size of the term/document matrix. The first column of Table 7 gives the number of terms after removing all stoplist words and conflating words via the stemming step. The second column shows the number of terms that are left after rejecting those that do not occur at least three times across the entire chronicon. And the penultimate column gives the ratio of the first two columns: the number of terms rejected for each term that is preserved. While the two NetNews chronica have a similar rejection rate (about 1.5 terms rejected for each one kept) the Globe chronicon has a noticeably higher rejection rate. In the next section a collection of stylostatistical measures will be used to explore this difference more closely. The last column is simply the number of terms multiplied by the number of documents: in other words, the size of the term/document matrix. The large sizes for these matrices contribute to the computational complexities of the clustering and analysis process.

I employed the Nearest Neighbor clustering algorithm along with a fixed radius to group the texts within each chronica topically. Table 8 gives the summarizing statistics for the sizes of these clusters. Clearly the Skeptic chronicon enjoys noticeably smaller clusters, probably due to the relatively broad range of topics discussed across this collection. The Clinton chronicon, in contrast, has a larger mean cluster size and a much larger standard deviation about this mean. The size of the clusters formed for this collection varies considerably. Finally, note that the largest cluster

within both the Globe and Clinton collections is roughly five times the size of the largest Skeptic cluster.

Chronicon name	Number of clusters	Mean cluster size	Standard deviation of cluster size	Max cluster size
Globe	4619	22.6	4.87	934
Clinton	1010	15.5	55.51	1139
Skeptic	3023	3.9	12.56	232

TABLE 8. Summarizing statistics for cluster sizes of three chronica.

Figure 24 shows the cluster size data of Table 8 on a semilog plot. I have sorted the cluster sizes for each text and plotted this sorted series of numbers for each chronicon. Note that the plots visually vary most in their start positions on the x-axis: the more to the right a plot sits, the larger the number of very small clusters. The Globe chronicon clearly contains many clusters of only a single text. Further, the wider the plots, the more even the distribution between large and small clusters. Notice, however, that the general shape of each of the plots is comparable suggesting that the overall structures described by the cloud of texts within this conceptual space are similar among the three chronica. Pocklington and Best (1999) have explored in greater detail the structures formed by these points in space and argue that their

properties are due to the conflicting pressures of stabilizing and disruptive selection.

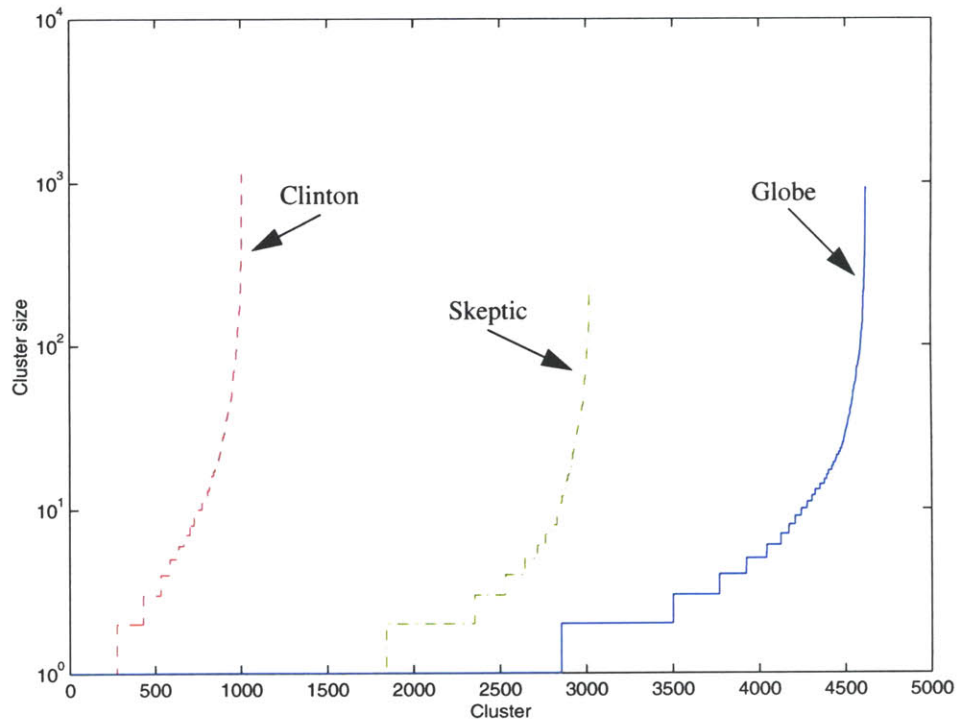


FIGURE 24. Clusters sorted by their size and plotted on a semilog graph. All three chronica show similar size distributions, though they vary most prominently in number of small clusters.

The results described in Chapter 4 (and the example results already discussed in Chapter 2) rely heavily on the largest clusters within each chronicon. These large clusters represent topics within the collections that have enjoyed significant treatment. I have closely inspected the texts within these large clusters, and in Table 9, I glossed each of these sets of texts. The clusters from the Globe chronicon form the most coherent collection of texts. This comes as no surprise as these texts were written by a smaller core of authors under some professional direction. The clusters from the two NetNews collections are a bit broader in their topics but, nonetheless, do describe coherent sets of concepts. Often these clusters represent the related

texts of one or a few large in-reply-to threads. For instance, the “Sex, sex, and more

Size	Gloss	Shorthand title
Globe clusters		
934	Clinton/Lewinsky scandal	Clinton/Lewinsky
545	Palestine/Israel conflict	Palestine/Israel
343	Northern Ireland/UK conflict	Ireland/UK
316	Conflict with Iraq	Iraq
293	Former Yugoslavia	Yugoslavia
285	Corrections to the Globe	Corrections
Clinton clusters		
1139	Sex, sex, and more sex	Sex&sex
743	Senate acquittal of Clinton	Senate
497	Interview with Linda Tripp	Tripp
363	Gun control	Guns
320	Juanita Broaderick acquisitions	Broaderick
304	Rants about Ted Turner	Turner
Skeptic clusters		
232	John Smith is a Nazi	Nazi
217	Unabomber discussion	Unabomber
207	McDonalds coffee suit	McDonalds
195	Solar energy	Solar
132	Language acquisition	Language
130	Military warheads	Military

TABLE 9. The six largest clusters from each of the chronica with a subject gloss and short title.

sex” cluster from the Clinton collection represents a wide range of posts offering a spirited and often smutty review of the Clinton/Lewinsky matter. Whereas, the “John Smith is a Nazi” cluster represents a prolonged debate about a particular person and whether or not his online behavior was inappropriate.

3.13 Stylostatistical Features for Two Chronica

The results outlined in the previous section offer certain clues to the nature of the three chronica. The variety of subject matter, language use, voice, and style all contribute to the size of the term list, the cluster sizes, and so forth. The corpuslinguistic community has developed a range of measures aimed specifically at quantifying aspects of text voice, genre, variety, etc. (e.g., Sinclair, 1991; Biber, 1993; Barnbrook, 1996; Liiv, 1997; Oakes, 1998). These measures are generally sensitive to the methods used to lemmatize the texts, which is in fact a very subjective business (Sinclair, 1991). The Porter stemmer was used for the Skeptic chronicon and for the Globe and Clinton chronica ENGTWOL was used. Thus, for my brief foray into stylostatistical measures, I will compare the Globe with the Clinton collection as they both were stemmed with the same system. Importantly, this will allow a comparison of internet discourse with printed media.

The key to understanding these stylostatistical measures is to appreciate the difference between *types* and *tokens*. A token is some language unit (in our case a lemma, but it could be just a letter) considered across a text and counting repetition. A type, in contrast, is the same language unit but without sensitivity to repetition. Consider the sentence, “I love my cat but otherwise hate cats.” If we lemmatize this sentence with ENGTWOL we arrive at the string, “I like I cat but other cat I hate.” Counting up instances, the token “cat” occurs twice, “I” occurs three times, and “other” occurs once. However, the type “cat” occurs only once since I do not count repetitions.

Liiv (1997) explored ten different stylostatistical measures from the corpuslinguistic community and used principal component analysis and other clustering approaches to distill just three measures that accounted for all the other. In other words, these three features characterized the texts just as well as all ten had when taken together. Consider a particular text from a collection; let N equal the number of tokens in the text, V equal the number of types, V_1 equal the number of words that occur only once within the text, and F_1 equal the absolute number of times that the most frequent word occurs within the text. Note that these values are computed for each *single* text and then averaged across all texts in a collection. Liiv’s three measures are: (1) N/V , the type-token ratio or mean word frequency which measures

the uniformity of the text. A large number would signify a text that heavily repeats the same words. (2) V_1/N , the index of rare words characterizes the variety or richness of the vocabulary by measuring the frequency of words within the text that only occur once. And, 3) F_1/N , the presence of the most frequent word, measures the concentration of the most frequent word in a text. Table 10 shows the mean values for each of these measures for all the texts within the two chronica, along with the standard deviation about this mean (in parenthesis).

Chronicon name	N/V uniformity	V_1/N variety	F_1/N concentration
Globe	1.159 (.054)	0.787 (.069)	0.091 (.030)
Clinton	1.146 (.067)	0.785 (.103)	0.085 (.173)

TABLE 10. Three discriminating stylostatistical measures for the Globe and Clinton chronica (Liiv, 1997). Mean value and standard deviation (in parenthesis) are shown.

Examining Table 10 it is clear that there are not significant differences in the mean values for these two chronica. The type-token ratio implies that each type appears on average about 1.2 times in a text. This is actually a fairly small value compared to many other collections (Barnbrook, 1996; Liiv, 1997; Oakes, 1998) and may be due to relatively small text sizes. The measure of variety, however, is similar across chronica and relatively high (Liiv, 1997; Gerbig, 1997). More than three-fourths of the texts on average are composed of words that only appear once. The standard deviation for the Clinton collection is noticeably higher than the Globe collection, suggesting more variance across texts in their variety of word usage. Finally, the measure of concentration, F_1/N , gives fairly standard values (Liiv, 1997) though again the variance is considerably higher for the Clinton chronicon.

These three measures describe average intra-document features across a collection of documents. But the difference in standard deviations across the chronica suggests that inter-document features, features that consider the set of documents together, may be worthy of study.

To briefly explore this, I have examined the hapax legomena across both of these chronica. Hapax legomena are words that are new to a collection or linguistic record. Thus, all words at some point are hapax, given some fixed set of texts; but, for common words, this occurs early on. Generally, half of all words within some

text are hapax legomena and these words are often the most interesting and delicate (Gerbig, 1997; Oakes, 1998).

Figure 25 shows the number of hapax legomena found in each text plotted against time (top graph) and against the cumulative count of tokens (bottom graph). Both of these graphs are quite similar, demonstrating fairly uniform text sizes. The appearance of these graphs seems quite standard (Baayen & Renouf, 1996). They show an early accumulation of hapax, as common words are first discovered, and a long tail representing the arrival of 10-200 new hapax for each text. This is a standard level of variety and linguistic production.

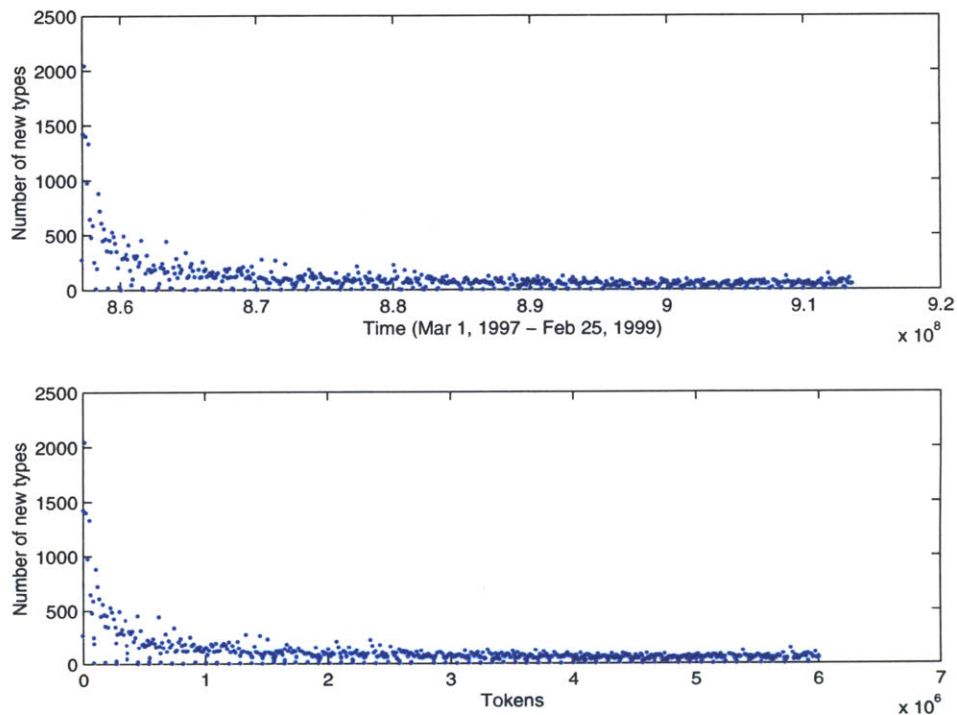


FIGURE 25. Hapax legomena for Globe chronicon plotted against time (top) and number of tokens (bottom).

Figure 26 shows the same two graphs for the Clinton chronicon. What these graphs indicate is significantly less hapax to start with and a much more narrow tail with less than 100 new words for each text. This suggests that the NetNews collection

has less lexical variety, less linguistic production, and so forth. This is, for me, a bit of a surprise as I expected that the free form nature and large number of authors in the NetNews collection would increase the hapax. For instance, I've experienced far more nonce terms, misspellings, and the like within the Clinton chronicon compared to the Globe chronicon. I now suspect that perhaps the Globe chronicon contains more sublanguages (Kettridge & Lehrberger, 1982) due to its often specialized or technical coverage; this should contribute to the generation of hapax legomena.

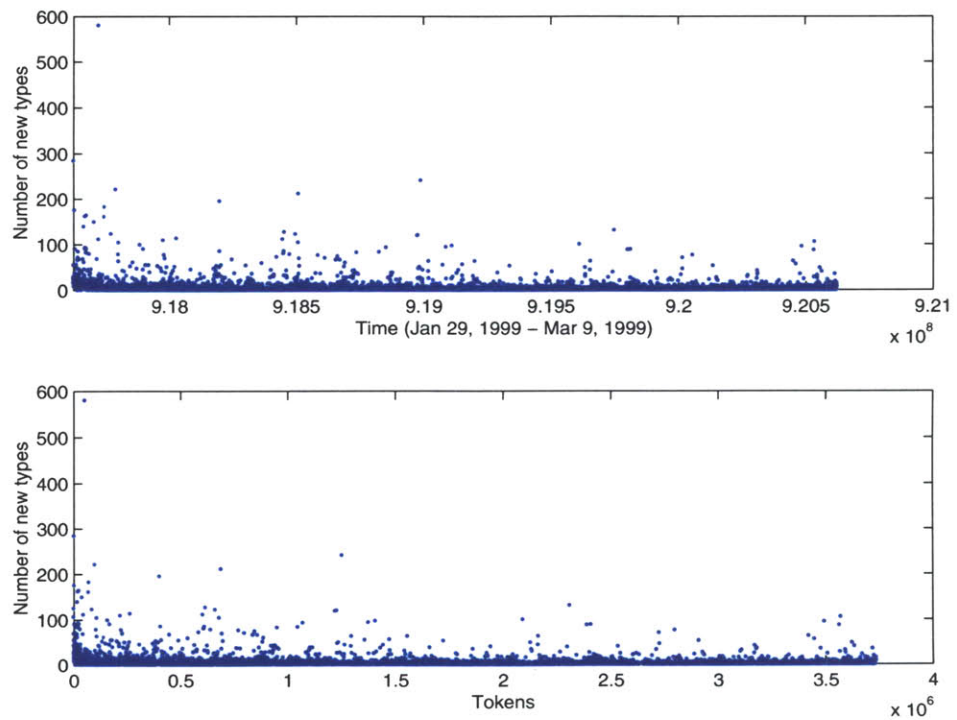


FIGURE 26. Hapax legomena for Clinton chronicon plotted against time (top) and number of tokens (bottom).

To my knowledge, this amounts to the first stylististical analysis of NetNews discourse. The measures reported here just give a flavor to the type of comparative corpus-linguistic research that can be carried out. Texts on the net are a fundamentally new and important means of communications, and I believe that genre analysis of these discussions is an exciting, open research area.

3.14 Summary

I described the core systems of the CAMEL text analysis package. The CAMEL system analyzes text collections, computes term vectors, clusters the texts topically based on these vectors, and so forth. It employs a variety of approaches to minimize the size of the term list including stemming or lemmatizing words to their base form.

I have developed a method to distill active replicators within the text collection from the large number of linguistic replicators identified by the CAMEL system. It is these active replicators that offer evidence of evolutionary activity and thus are most fundamental to the microevolutionary model. Active replicators are identified via the Pearson product-moment correlation coefficient. Computing correlations between timeseries is fraught with peril, so I have spent a considerable amount of time considering potential sources of statistical artifact, including problems with trend, drift, and autocorrelation. In the end, however, it is a qualitative assessment of the results that is the most powerful confidence builder.

I ended the chapter with a brief foray into comparative corpuslinguistics. A few discriminating stylostatistical measures were computed for the Globe and Clinton chronica. This demonstrates the sort of analysis that can be fruitfully applied to internet text collections.

Active Language Replicators

This chapter reviews data that form the core results in support of the Microevolutionary Language Theory. I use the CAMEL system and chronica described in Chapter 3 to distill linguistic replicators. These replicators are at multiple levels, namely, lexical (words), lexical co-occurrence (groups of words), lexico-syntactic (words in their structural role), and syntactic (just the structure). At each of these levels (save lexical) I need to make use of additional software mechanisms, so my presentation of results is intermixed with expositions on these additional software systems. I studied lexical, lexico-syntactic, and syntactic replicators within the Globe and Clinton chronica and lexical co-occurrence within the Skeptic chronicon.

The punch line goes something like this: I was able to find active replicators, units of replication that are autocatalytic, at all linguistic levels save the syntactic level. At the syntactic level I found one or a few replicators that *may* be active, but I was not able to demonstrate this to my satisfaction.

4.1 The Centrality of Level

Since the 1940's, the notion of level has been central to the linguistic programme. In fact, linguistics of the latter 20th Century has been primarily engaged in studying how a *langue*, a language system as a static social reality, exists at various untangled and pure levels (phonetic, lexical, syntactic, and so on).

The level of linguistic analysis in vogue amongst researchers has undergone significant change over time. The early lexicologists considered the word most worthy of study (Gerbig, 1997). The structuralists relegated lexis to the trash heap, considering it a trivial irregularity of language not worth their time (Martinet, 1960/1975). Even today linguists are prone to dismiss lexis. Consider Lass, who is often good for a quote one way or the other: “There’s something curiously amateurish and pop-linguistic about a pre-occupation with lexis at the expense of structural relations” (1997, p. 169). And even if the generative school saw itself as apart, it took a page from the structuralists in its dismissal of lexis and concentrated wholly on syntax (Chomsky, 1965). And so it goes.

Recent significant results from the corpuslinguistic community, however, show that a clean split between lexis and syntax is not possible insofar as it is not maintained by human language in use. Language exists in its relation of form with meaning; ultimately there is no distinction between them (Sinclair, 1991; Györi, 1995). As Gerbig nicely put it: “Corpus linguistics has demonstrated that lexis and syntax reciprocally determine each other and are therefore co-selected” (1997, p. 97). The patterning of words is not independent of the words themselves. Thus a new level of analysis is demanded — that of lexico-syntax, words and their structural relation.

Linguistics is not the only discipline obsessed with the question of level. Evolutionary theory has taken a turn or two around the same problem. In the next chapter I will take up the issue of appropriate units within evolving systems (see Section 5.8). Suffice it to say that in evolving systems, including biological ones, selection may simultaneously favor multiple units at different and interacting levels of selection (e.g., genes, individuals, kin groups) (see in particular Lewontin, 1970; see also Sober, 1984; Breden & Wade, 1989; Breden & Hausfater, 1990)

Thus, an answer to the question of what levels of language evolve may well be: yes. Evolution may occur at multiple levels. To explore this, I will examine microevolution at multiple levels of language: lexical, lexical co-occurrence, lexico-syntactic, and syntactic. And my results support a theory that (micro-)evolution occurs and is observable at multiple linguistic levels.

4.2 Lexical Replicators

The simplest level to be examined is lexical. That is, I explore the microevolutionary dynamics of individual words. Both the *Globe* and the *Clinton chronica* were studied at this level of analysis. Happily, studying lexical replicators is relatively straight forward. Armed with the methods described in Chapter 3 and an under-

standing of the distinction between a replicator and an active replicator I am prepared to plunge into the results.

The CAMEL system lemmatizes words with ENGTWOL and counts the number of occurrences of the lemmas across the texts. The results of the previous chapter show clearly that some terms frequently reoccur throughout the text, and some terms occur but once or perhaps a very few times (“terms” and “lexemes” will be used interchangeably to distinguish the set of stem-words from the set of all words). Lexical replicators are terms that occur with reasonable frequency. In Table 11, I have listed the top replicating terms throughout the entire Globe and Clinton chronica having for this table returned the stop words to the mix. Two things should be clear: First, these terms occur many, many times within these texts. Second, they are all grammatical or function words. That is, they only have grammatical meaning (e.g., constructions, inflections, etc.) in contrast to lexical meaning. I would not expect these sort of words to enjoy active replication on the timescale of these chronica, and indeed, find that they do not. Thus, even though these words enjoy very high levels of replication, within the short timescales resolved by these text collections they are not under observed evolutionary pressures.

Globe chronicon		Clinton chronicon	
Lexeme	Count	Lexeme	Count
the	413,892	the	222,518
of	201,848	to	122,094
to	197,954	of	104,700
a	175,144	and	93,352
and	161,133	a	87,734
in	147,587	that	69,355

TABLE 11. Top six lexical replicators from Globe and Clinton chronica (stop words considered). Not surprisingly, they are all function words.

4.2.1 Globe active replicators

Having considered simple lexical replicators let's turn to the question of active replicators and launch straight into results from the Globe chronicon. I searched for active replicators by examining lexemes whose relative presence correlates strongly with the publication volume within populations of texts (Section 3.10). These text populations are the topic clusters described in Table 9. An active replicator is any lexeme that strongly correlates with a cluster's volume of texts. Listed in Table 12 are the most active replicators from the cluster of texts dealing with the Clinton/Lewinsky scandal for those lexemes that occur at least 40 times. For the words, I have chosen to consider a replicator active if $r \geq 0.5$ and $n \geq 40$. Many of these active lexical replicators will be recognized from Table 3 of Chapter 3.

<i>r</i>	<i>p</i>	<i>n</i>	Lexeme
0.6333	< 0.000001	51	moral
0.5866	< 0.000001	408	public
0.5537	< 0.000001	60	denial
0.5391	< 0.000001	768	president
0.5334	< 0.000001	73	swear
0.5270	< 0.000001	309	lewinsky
0.5255	< 0.000001	61	true
0.5178	< 0.000001	128	word
0.5170	< 0.000001	86	behavior
0.5165	< 0.000001	307	affair
0.5008	< 0.000001	759	Clinton
0.5008	< 0.000001	331	sex

TABLE 12. Active replicators from Clinton/Lewinsky cluster (stop words excluded). Cluster size is 934 texts.

Consider for a moment the strongest active replicator for this cluster, namely, the word "moral." It bears repeating what this correlation suggests: the more a text reporting on the Clinton/Lewinsky scandal repeats the word "moral," the more texts

are published on the Clinton/Lewinsky scandal. Recall that this represents the stem-word; therefore, “morality,” “morals,” “moralizing,” etc. may all also be occurring. In Figure 27, I have plotted the two timeseries associated with this active replicator. The solid line describes the number of texts published week-by-week within this cluster. The dashed line describes the relative presence of the lexical replicator “moral.” The covariance between the two timeseries is clear.

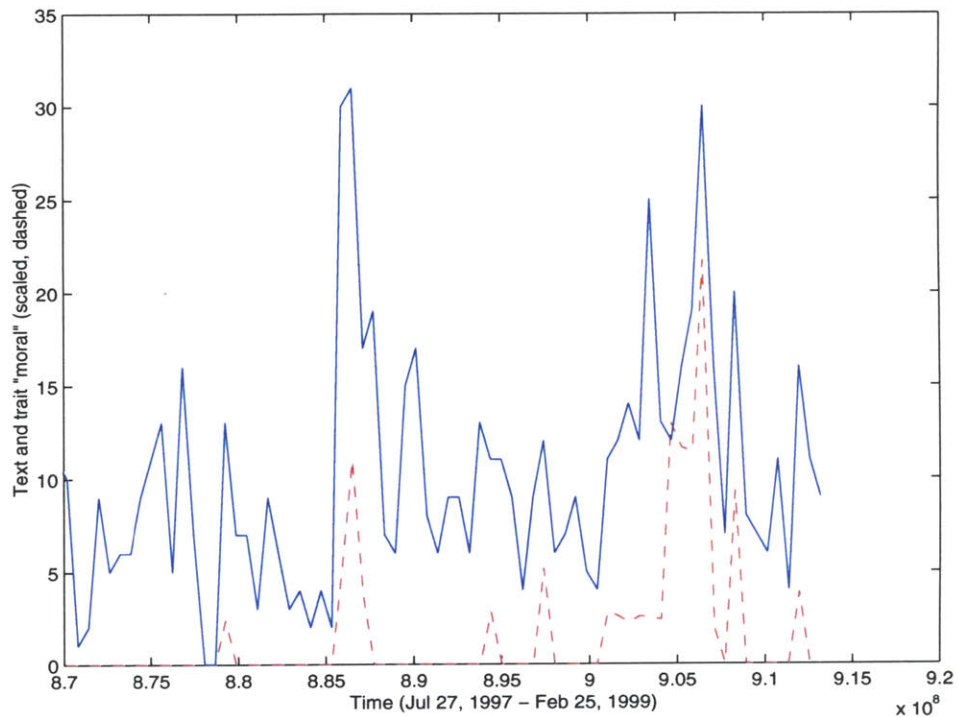


FIGURE 27. Lexical replicator “moral” (dashed) and number of texts in a week published within the Clinton/Lewinsky cluster of the Globe chronicon (solid, $r = 0.6333$).

I computed such correlation coefficients for all common replicators against all large clusters. Given the criteria I set forward (namely, $r \geq 0.5$ and $n \geq 40$) I discovered sets of active lexical replicators for two other large clusters. The cluster of texts associated with the U.S. war in Iraq included a large number of active replicators. The top fifteen are recorded in Table 13; the top few are highly correlated with text volume.

Active Language Replicators

<i>r</i>	<i>p</i>	<i>n</i>	Lexeme
0.7264	< 0.000001	221	US
0.7154	< 0.000001	189	Hussein
0.7145	< 0.000001	77	house
0.6998	< 0.000001	164	relation
0.6805	< 0.000001	189	Saddam
0.6592	< 0.000001	96	strike
0.6562	< 0.000001	187	president
0.6527	< 0.000001	61	senior
0.6519	< 0.000001	67	send
0.6434	< 0.000001	273	yesterday
0.6309	< 0.000001	61	white
0.6158	< 0.000001	81	diplomatic
0.6002	< 0.000001	119	Clinton
0.5986	< 0.000001	294	Iraq
0.5950	< 0.000001	143	force

TABLE 13. Fifteen top active replicators from the conflict with Iraq cluster of Globe chronicon (stop words excluded).

In Figure 28, I plot the relative presence of the “Hussein” lexical replicator against

the volume of texts published within the Iraq cluster. The correlation is striking.

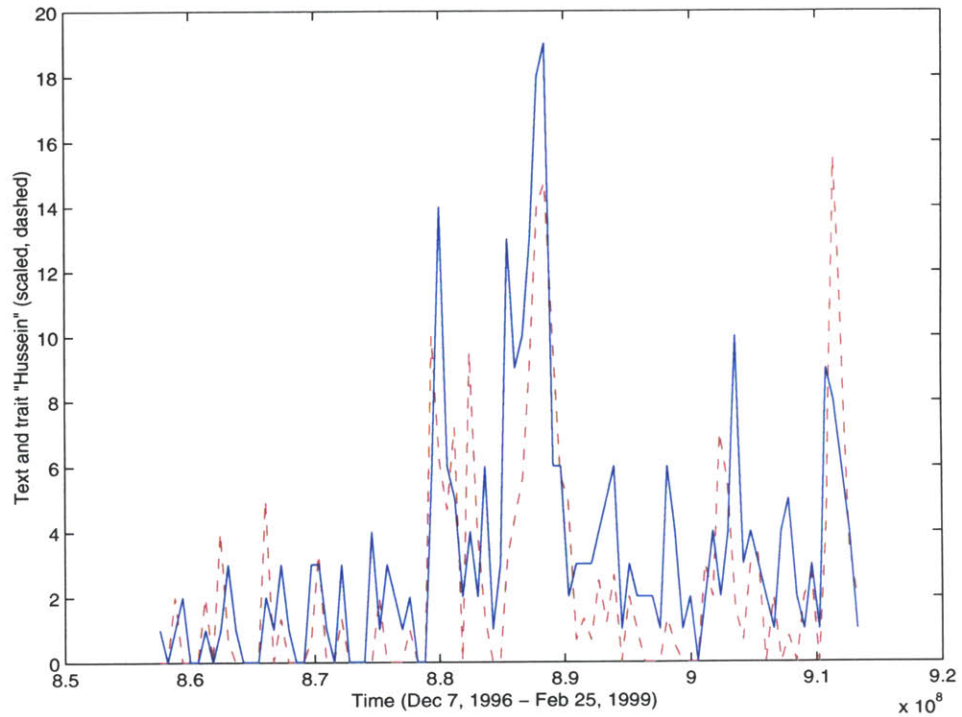


FIGURE 28. Lexical replicator “Hussein” (dashed) and volume of texts within the Iraq cluster in the Globe chronicon (solid, $r = 0.7154$).

The second cluster that displayed active replicators was not listed in Table 9. It is a cluster of texts which dealt with the U.S. Embassy bombing in Kenya (given the shorthand title “Kenya”). The cluster contains 212 texts. The top five active replicators are given in Table 14.

Active Language Replicators

<i>r</i>	<i>p</i>	<i>n</i>	Lexeme
0.6587	< 0.000001	91	attack
0.6365	< 0.000001	125	US
0.5351	< 0.000001	52	bombing
0.5281	< 0.000001	179	yesterday
0.5118	< 0.000001	85	embassy

TABLE 14. Top five active replicators from the Kenya cluster of the Globe chronicon (stop words excluded).

In Figure 29, I plot the timeseries associated with the “attack” lexical replicator against the timeseries associated with the cluster of texts. Again, the correlation between these two series is striking. (Note that I am plotting the timeseries for lexical replicators to give a feel for what a high correlation coefficient corresponds to visually. In subsequent sections, I will dispense with this practice, as the values for

r , n , and p should suffice.)

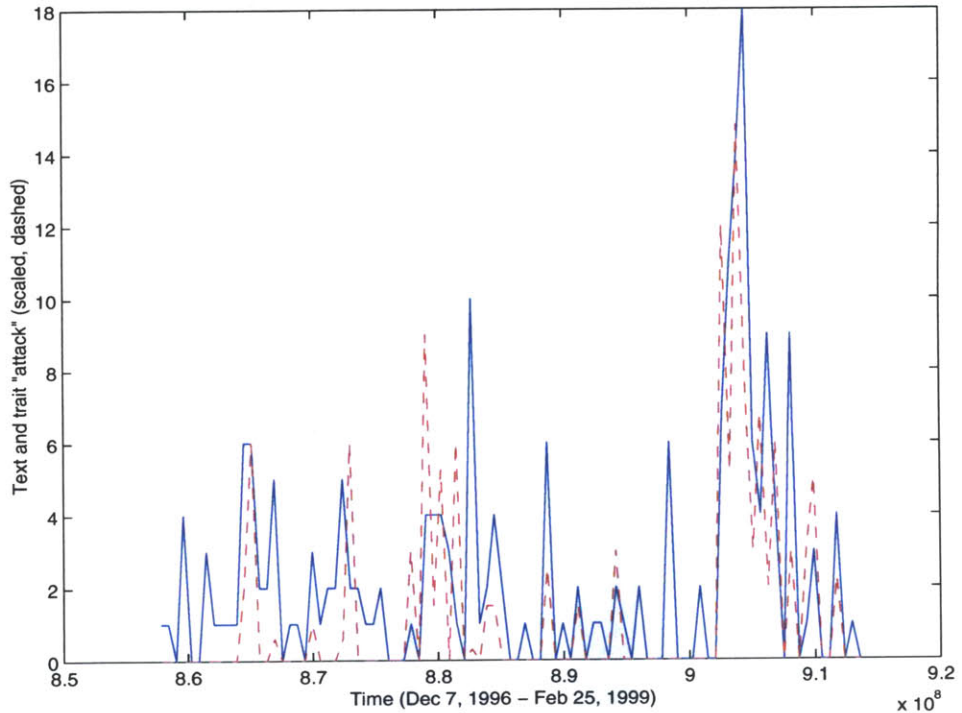


FIGURE 29. Lexical replicator “attack” (scaled, dashed) and volume of texts associated with the Kenya cluster within the Globe chronicon (solid, $r = 0.6587$).

4.2.2 Summary and analysis

I examined the eleven clusters from the Globe chronicon that are larger than 200 texts. Amongst them, 53 replicators which appeared at least 40 times had correlation coefficients of 0.5 or greater, I consider these to be active lexical replicators. (The average correlation coefficient for a lexical replicator within these texts was -0.0771.) These replicators occurred across six clusters but the bulk of them were from the Iraq cluster. Figure 30 shows the number of active replicators for each of

the ten large clusters. In the previous section, the most correlated lexemes for the three clusters with a large number of active replicators were described.

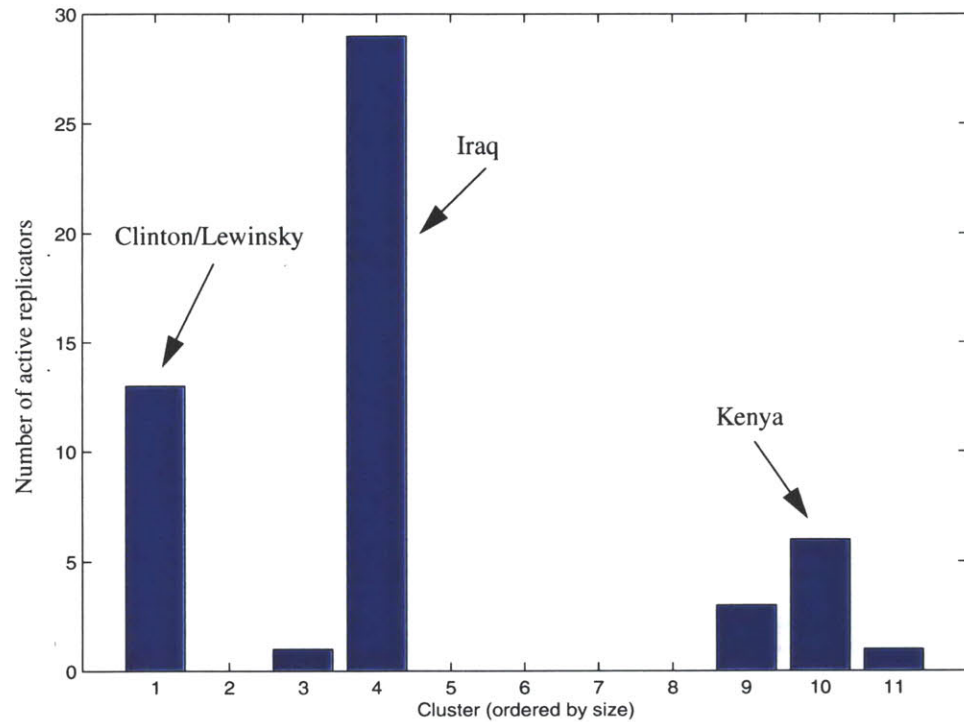


FIGURE 30. Number of active lexical replicators for the ten clusters within the Globe chronicon with more than 200 texts.

For these three clusters the active replicators certainly are evocative. Besides a list of the major players in each story (the President, Monica Lewinsky, Saddam Hussein, etc.) they also consist of other words that are at the center of their cluster's theme. These are powerful and compelling words. For the Clinton/Lewinsky stories we see matters of truth, morals, lies, and sex. For the Kenya and Iraq topics, there are words like attack, force, strike, and bombing. Interestingly, "yesterday" occurs for both of these clusters. In fact, "yesterday" is an active replicator for four of the top eleven clusters. This, I suspect, is the result of a daily newspaper's need to be timely — very current stories are always the hottest commodity. I did not compute the correlation coefficients for the stop words; but the fact that the active lexical replicators all strongly bear content suggests this did not alter the results.

Why do the active replicators clump together in three clusters leaving many clusters with only relatively weak correlations to their terms? Is it something about these clusters and their themes and topics? Is it a matter of chance and contingency? So far I do not have an answer to these questions.

4.2.3 Clinton active replicators

As mentioned earlier, the NetNews collections do not produce clusters centered on as clear-cut topics as are seen within the Globe newspaper collection. Indeed, they are somewhat less satisfying and I will spend less time with the results from these chronica. Within the Clinton chronicon there are ten clusters with 200 or more texts. The distribution of active replicators across these ten clusters is shown in Figure 31.

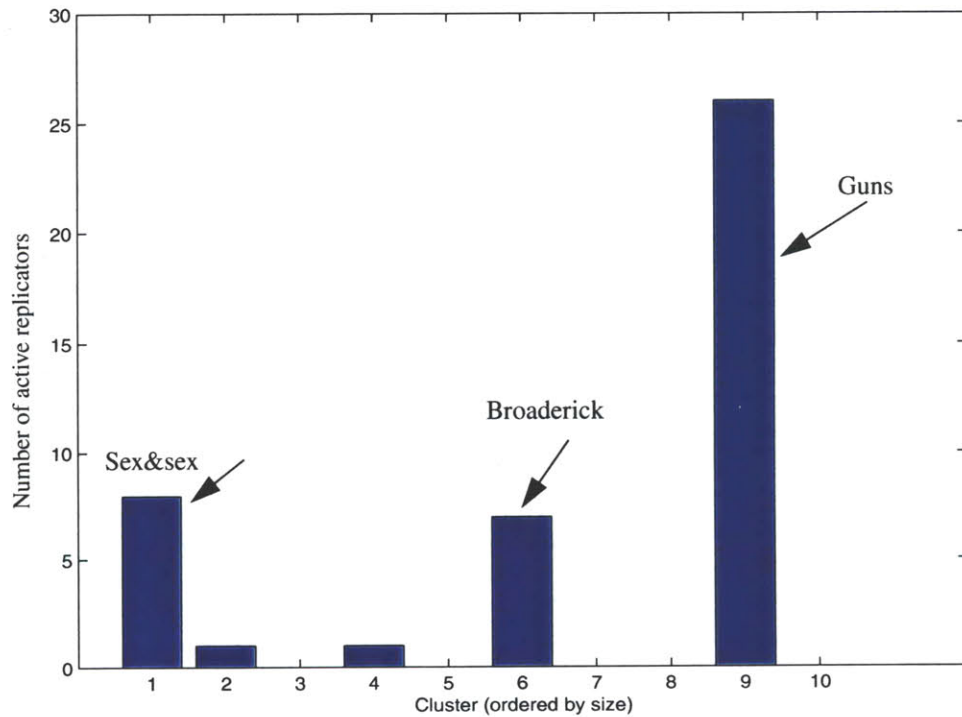


FIGURE 31. Number of active lexical replicators for the ten clusters within the Clinton chronicon with more than 200 texts.

Active Language Replicators

Cluster title	<i>r</i>	<i>p</i>	<i>n</i>	Lexeme
Sex&sex	0.6671	< 0.000001	928	write
	0.6550	< 0.000001	311	Feb
	0.5810	< 0.000001	91	bastard
	0.5425	< 0.000001	100	Jim
	0.5258	< 0.000001	94	sexual
Broaderick	0.6001	< 0.000001	309	Re
	0.5637	< 0.000001	186	lie
	0.5528	< 0.000001	116	broaderick
	0.5440	< 0.000001	104	cunt
	0.5420	< 0.000001	51	guilty

TABLE 15. Top five active replicators for two of the clusters from the Clinton chronicon.

This list of active replicators (Table 15) shows some of the problems with an analysis at the lexical level. For instance, a very strongly correlated lexical replicator for the Broaderick cluster is “Re.” Recall that the string “Re,” short for “Regarding,” occurs quite often on the subject line of in-reply-to posts within NetNews (see Figure 12). The four other active replicators listed for this cluster (“lie,” “broaderick,” “cunt,” and “guilty”) have clear semantic relevance to these texts. But the “Re” replicator is not lexically related to the texts, instead it serves a structural role that calls out that this text is an in-reply-to a previous post. Why does it maintain such a strong correlation with the volume of texts (why is it *active* at all)? I believe it is because it co-occurs with other words that do have strong semantic currency. In other words, there are larger replicating complexes that are not resolved at the lexical level.

Upon a closer examination of the texts in this cluster, I see that there are a number of subject lines where “Re” collocates with the other active replicators shown in Table 15, including “Re: Juanita Broaderick is a lying cunt” and, for that matter, “Re: Juanita Broaderick is a brave soul.” Thus, “Re:” amounts to a *spandrel* (Gould & Lewontin, 1979; but cf. Dennett, 1995); it is a structural tag-along, an outcome of

the physical environment within the NetNews system, that associates with some other adaptive component. (In Section 5.8 I will discuss these issues further.) In the examination of the lexical co-occurrence and lexico-syntactic level some of these collocations will form a complete unit of analysis.

4.3 Lexical Co-occurrence

In the previous section I described a lexical replicator, “Re,” that has a strong autocatalytic correlation. But I suspect this is due to the company it keeps more than to any quality in its own right. One way to explore this issue is to examine the co-occurrences within a text and look for active replication amongst sets of words occurring together.

In order to study this level of replication I’ve developed additional software mechanisms based on a technique called Latent Semantic Indexing (Furnas, et al., 1988; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Dumais, 1992; Dumais, 1993). This technique identifies sets of terms that co-occur together with statistical frequency across a collection of texts. Note that my studies of replicating co-occurring terms were performed on the Skeptic chronicon, instead of the Globe or Clinton collection, and employed the Porter stemmer, instead of the ENGTWOL lexical analyzer.

4.3.1 Latent Semantic Indexing

In Chapter 3, I described a set of steps which culminated in the term/document matrix of Figure 14. In order to distill frequently *co-occurring* sets of terms I’ll take an additional step and attempt to discover higher-order structure within the matrix. That is, I’ll find the statistically salient associative relationships caused by term co-occurrence. This is done through a principal component analysis called singular value decomposition or SVD.

Matrix decomposition techniques, such as SVD, are employed generally for two purposes — data reduction (compression) and data interpretation. Since the chronica results in very large and quite sparse term/document matrices (upwards of tens-of-thousands of texts by tens-of-thousands of terms, see Table 7) it is useful to perform some data compression insuring that the continued analysis remains computationally tractable. But more importantly, I will make use of the salient conceptual structures present in the term/document matrix to study replicating co-occurrences of terms. Thus, applying SVD to the term/document matrix both compresses the data and distills out the salient underlying semantic structures.

The use of SVD for text-retrieval applications was originally proposed and has been extensively studied by Susan Dumais, of Bell Communications Research, and her colleagues (Furnas, et al., 1988; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Dumais, 1992; Dumais, 1993). They refer to this technique as Latent Semantic Indexing (LSI). Peter Foltz has investigated the use of LSI in clustering NetNews articles for information filtering and text retrieval applications (Foltz, 1990). He studied small collections of posts from three newsgroups (comp.windows.x, soc.women, rec.ham-radio) and one bulletin board system. Michael Berry and co-authors have researched a variety of numerical approaches to efficiently perform SVD on large, sparse matrices such as those found in text retrieval applications (Berry, 1992; Berry, Do, O'Brien, Krishna & Varadhan, 1993; Berry & Fierro, 1995).

The singular value decomposition is formulated as

$$A_k = U\Sigma V^T = \sum_{i=1}^k u_i \cdot \sigma_i \cdot v_i^T,$$

where A_k is a rank- k approximation to A . That is, the matrix A is decomposed into k left and right eigenvectors (u_i and v_i) and a diagonal matrix (σ_i) composed of the k eigenvalues (Figure 32). The SVD approach insures that A_k will be a good approximation, in the least-square error sense, to the original A . *But it should not be exact*, since this process is used to remove noise from the original matrix while keeping the most salient data. The structure in the term/document matrix is used to re-express its data in a more parsimonious fashion — one in which this structure is brought to the surface unobstructed.

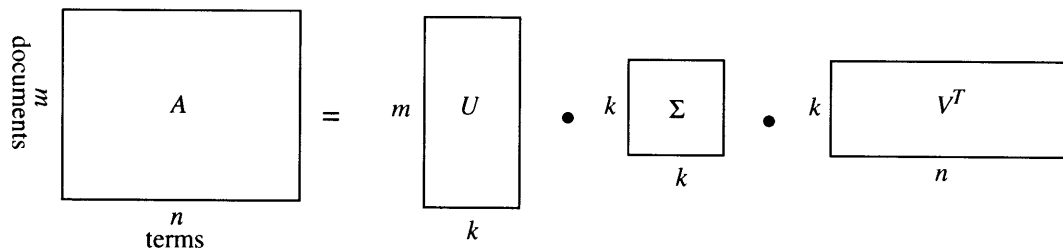


FIGURE 32. Decomposition of term/document matrix into rank- k approximation.

Only a relatively minuscule percentage of all of the terms in a chronica appear in any particular document. Therefore, the term/document matrix should be extremely sparse. Moreover, the matrix will have a fair amount of underlying structure due to word co-occurrences within the posts. More formally, the term/document matrix will have a very low rank relative to its dimensions, $k < r \ll \min(m, n)$. In practice, the dimensions of A are generally on the order of tens- or hundreds-of-thousands while the rank is on the order of hundreds. The SVD is, therefore, *low-rank revealing* and these top k vectors will provide a good approximation of the fundamental numerical subspaces present in matrix A (Berry, 1992).

But what exactly do these numerical subspaces represent? They are vectors that define linear combinations of either terms (for the right matrix of eigenvectors) or documents (for the left matrix). We say that the right matrix defines “*term-subspaces*.” Each term-subspace describes a set of semantically significant associative patterns in the terms of the underlying collection of documents: that is, a set of co-occurring terms. Each subspace acts as a *conceptual index* into the corpus (Furnas, et al., 1988).

4.3.2 Example

A small collection of NetNews posts has been examined for test and explanatory purposes. The collection is comprised of 784 texts composed of 82 posts to sci.military.moderated, 490 to sci.military.navy, and 212 posts distributed among the groups sci.psychology, sci.psychology.theory, sci.psychology.personality, sci.psychology.psychotherapy, sci.psychology.journals.psychology, sci.psychology.misc, sci.psychology.announce, and sci.psychology.research. The posts were made during the month of September, 1995. Of all the words in the set of documents, 5162 terms made it through the stop-list and stemming process. Thus, the final term/document matrix was 784 x 5162 elements in size. Each document, on average, was composed of 74 terms, and each term on average appeared in 6 documents. The term/document matrix was input to the SVD software which decomposed the matrix into a rank-265 approximation: the SVD determined that the term/document matrix was best compressed into 265 subspaces. Considering just the right matrix, each column vector describes a set of associative relationships of words as a weighted linear combination of the original terms. Generally, each singular vector has only a small number of terms of significant weight. Thus, when describing each singular vector (or term-subspace of co-occurring words) I con-

sider only those terms with weights above some particular threshold, normally 0.1 (these vectors are unit normalized).

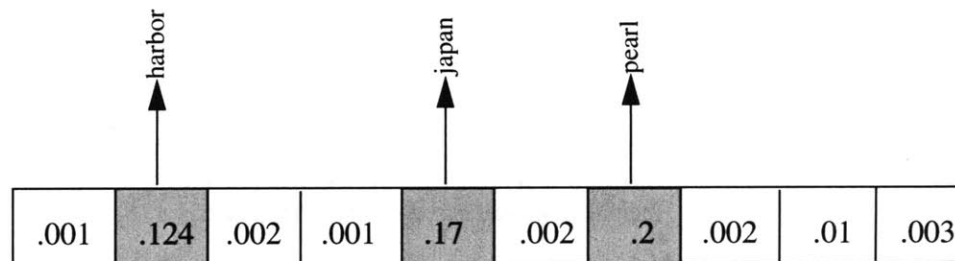


FIGURE 33. Most significant weights in the vector (shaded) represent the salient terms. Taken together they describe a set of co-occurring words that replicate together.

Now I will only consider the salient terms (with weights greater than 0.1) from five example term-subspaces of this collection. That is, I list those most significant terms of these singular vectors. Each singular vector or term-subspace describes a statistically significant set of term co-occurrences. Remember that these documents were either posted to military or psychology newsgroups:

- harbor, japan, pearl
- food, maze, rat, reinforce
- airforce, arsenal, tomahawk
- explode, meltdown, nuclear, russia, sub

From this example I found a set of lexical co-occurring replicators, words that co-occur together within a text and replicate together across the chronicon. Thus, this is a new type of replicator occurring at a different linguistic level. Am I able to find active replicators at this level?

4.3.3 Active lexical co-occurrence replicators

The Skeptic chronicon was studied using the SVD technique. And in particular I looked at the six largest clusters (only three clusters contained 200 or more texts). These clusters were determined using the approach detailed in Chapter 3 and did not make use of the term co-occurrences. The number of active replicators for these

six clusters is shown in Figure 34. Some of the replicators, listing the salient terms from each subspace, are shown in Table 16.

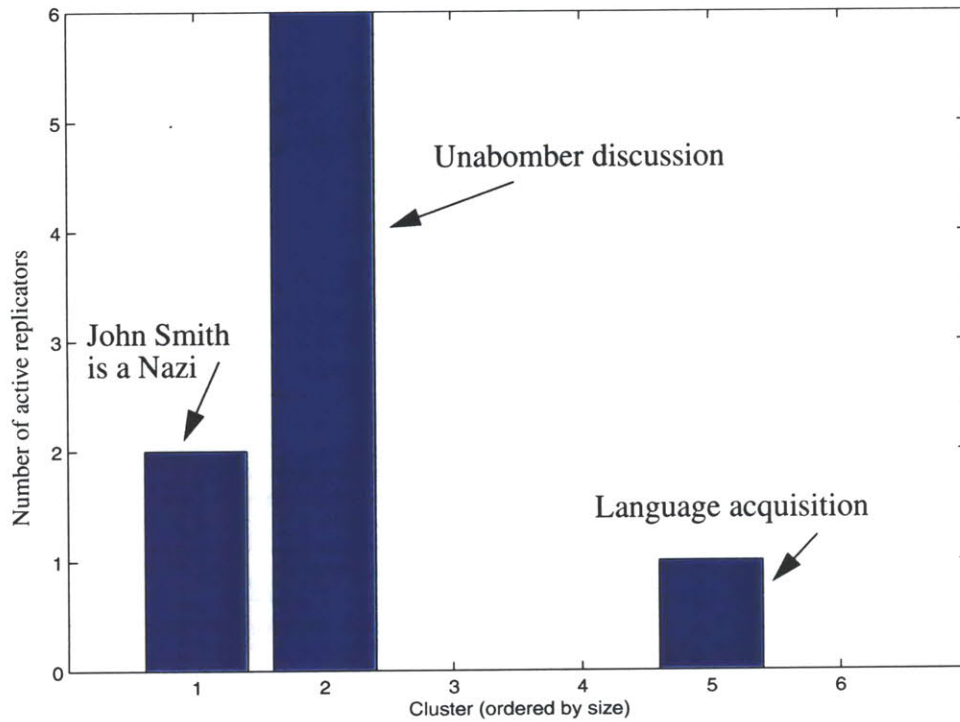


FIGURE 34. Number of active lexical co-occurrence replicators for the six largest clusters within the Skeptic chronicon.

I have examined at some length the “John Smith is a Nazi” cluster (Best, 1998a; Best, 1999a). The main thread of discussion within this particular set of texts originated with a collection of posts under the subject line “Homeopathy for Dummies” and “Homeopathy in TIME Magazine.” At first the discussion did focus on homeopathic remedies, but the 232 posts which populate this particular cluster deal with a flame-war centered around an individual poster who had been involved in the original discussion on homeopathy. This individual was apparently a wildly prolific author of posts and his style was considered by some to be confrontational and insulting. After time the cluster mutated its subject line so as to make more explicit its concentration on this individual. In fact, the largest number of posts in the cluster use the subject line “Is John Smith a Nazi?”¹. This subject line also mutated

Active Language Replicators

over time; interestingly, one such mutation read, “A Plea for Netiquette (was Re: Is John Smith a Nazi?)” Many of these texts were devoted to calling John Smith a Nazi, or to defending him from such attacks.

Within this cluster, an active replicator contained the co-occurring terms “John,” “Smith,” and “Nazi.” This replicator enjoyed considerable covariance with the volume of posts to the thread. In fact, this is the only lexical co-occurrence replicator within this cluster that had a high enough correlation coefficient to be considered active (see Table 16).

Cluster	<i>r</i>	<i>p</i>	<i>n</i>	Lexical co-occurrence
John Smith	0.8408	< 0.000001	101	John, Smith, Nazi
Unabomber	0.8460	< 0.000001	208	motive, science
	0.7510	< 0.000001	174	coffee, motive, science
	0.6480	< 0.000001	138	motive, science
	0.6370	< 0.000001	152	asthma, motive, science
	0.5742	< 0.000001	137	????

TABLE 16. Active replicators from two large clusters within the Skeptic chronicon.

The discovery of this particular replicator encouraged me to examine the use of “Nazi” as a pejorative attack word within other posts to NetNews. (Anyone who reads NetNews knows that “Nazi” is often used to flame people.) In a brief foray outside of my principal three chronica I easily found occasions of “Nazi” within the term subspaces for two other newsgroups. One cluster was from the soc.subculture.bondage-bdsm newsgroup (enthusiasts of sexual bondage and discipline) and the other was from alt.politics.usa.constitution (discussions on the U.S. Constitution). All of these replicators were examples of “Nazi” used as an attack word and not with reference to German National-Socialists (this was verified by inspection). And for both of these clusters of texts I was able to identify very strong correlations between the relative appearance of the lexical co-occurrences (“Nazi” along with the appropriate target of the attack) and the volume of posts to the thread. Table 17 summarizes these results for all three chronica, and the timeseries are delineated in

1. I use here a pseudonym.

Figure 35. Clearly, “Nazi” forms a part of a lexical co-occurrence replicator that is active across a variety of topic areas (scientific skepticism, sexual fetishism, and the U.S. Constitution).

Chronicon	Total number of texts	Dates	<i>r</i>	<i>n</i>
Skeptic	11,758	9/20/95 - 9/26/95	0.8408	101
soc.subculture.bondage-bdsm	1,160	9/28/97 - 10/6/97	0.8166	66
alt.politics.usa.constitution	494	10/30/97 - 11/2/97	0.5308	29

TABLE 17. Three clusters from differing NetNews newsgroups. The “Nazi” replicator has high correlation with volume of texts from three different topics.

Returning to the Skeptic chronicon and Table 16, a collection of active replicators from the “Unabomber discussion” cluster is also listed. This thread of texts center on a document authored by the Unabomber where he discusses cynically the motives of the scientific enterprise. A collection of highly correlated replicators contain the two terms “motive” and “science.” In general, these results are not as encouraging as the previous “Nazi” replicators. The LSI technique distills statistical co-occurrences that are not always understandable on the face of it. The replicator with only “motive” and “science” seems to make sense. But the addition of “coffee” and “asthma” seem likely to be artifacts and not particularly relevant to these texts. Finally, the lexical co-occurrence marked “????” did not have any salient terms but instead weakly represented a smattering of terms. The active replicators from other clusters are equally unsatisfying.

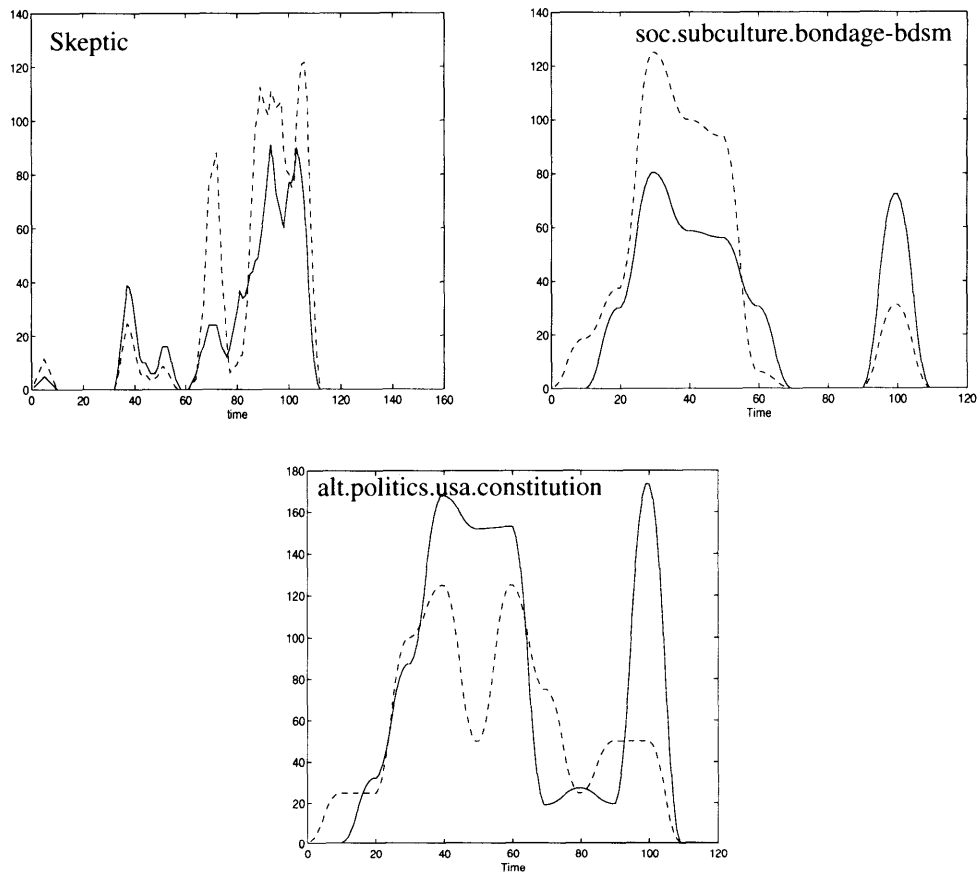


FIGURE 35. Lexical co-occurrence replicators with pejorative “Nazi” (dashed) along with volume of texts from Skeptic, soc.subculture.bondage-bdsm, and alt.politics.usa.constitution chronica (solid).

4.3.4 Summary and analysis

I applied the LSI method to distill sets of co-occurring words that reoccur with frequency across a collection of texts. This method was used with some success on the Skeptic chronicon in the discovery of a replicator where “Nazi” was used as a pejo-

rative attack word. I discovered this same type of replicator in three unrelated Net-News newsgroups. Identifying the same active replicator across multiple groups responding to similar selective pressures builds evidence that this trait is an adaptation. I will take this idea further in Chapter 5.

Beyond one significant success, the LSI technique did not perform very satisfactorily. Many active lexical co-occurrence replicators did not represent a set of semantically related co-occurring words. While I still maintain that word co-occurrences should resolve the problems of tracking spandrels as replicators at the lexical level, the SVD technique is clearly not a panacea. An open question is whether the LSI technique would perform materially better on a more stable chronicon, such as the Globe collection, and whether other simpler methods of distilling co-occurring terms would also perform better.

4.4 Lexico-syntax

Evidence was cited in the section above that words and their syntactic role are so reciprocally determined that it makes no sense to attempt to pull them apart (e.g., Sinclair, 1991; Gerbig, 1997). But in this dissertation I shamelessly have it both ways: in Section 4.2, I examined words removed from their structure and, in Section 4.5, I will talk about structure in isolation from any words. In this section I will practice what Sinclair and others preach (and what I, in the final analysis, believe) and examine words and their related structure held together. In order to accomplish this I need to extract a fair bit of structural information from a parse of the texts. Towards that aim I employ the Conexor Functional Dependency Grammar of English (FDG) (Conexor oy, 1998b), a close cousin to the EngCG-2 system containing the ENGTWOL lexical analyzer. My experiments on lexico-syntactic replication were performed with the Globe and Clinton chronica.

4.4.1 English FDG

The English FDG (Tapanainen & Järvinen, 1997) performs a surface-syntactic parse of free text. The parser establishes the head-modifier relations between words and develops a dependency tree. Within a phrase some word, the *head*, determines the syntactic range for that phrase, and this head may have a series of dependents, complements or *modifiers* that form arguments (Jackendoff, 1972; C.L. Baker, 1996; Radford, 1997). The FDG links words according to these dependency relationships but also labels these links with the syntactic function of the modifiers. Figure 36 shows a simple example of the dependency tree formed from the sen-

tence “I see a bird.” The head of each phrase is often a verb and the head of the entire sentence, marked with *main*, is the main verb.

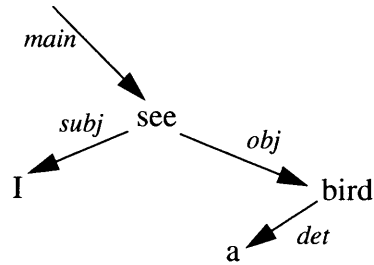


FIGURE 36. The dependency tree, with links labeled with their syntactic role, for the sentence “I see a bird” (adapted from Tapanainen & Järvinen, 1997).

The English FDG accomplishes this analysis by initially relying on the tokeniser, lexical analyzer, and morphological disambiguator of EngCG-2. It then employs a grammar of 2,500 rules that explicitly encodes the dependency relations. Having these relations explicitly in the grammar has been shown to aid considerably in disambiguating syntactic assignments (Tapanainen & Järvinen, 1997). Figure 37 shows a more complex example output from the FDG for the sentence “The theory here developed will be found to be based upon causality.” The syntactic tagset is

based on the original EngCG system (Karlsson, Voutilainen, Heikkilä & Anttila, 1995). The primary link tags are given in Table 18 (Conexor oy, 1998b).

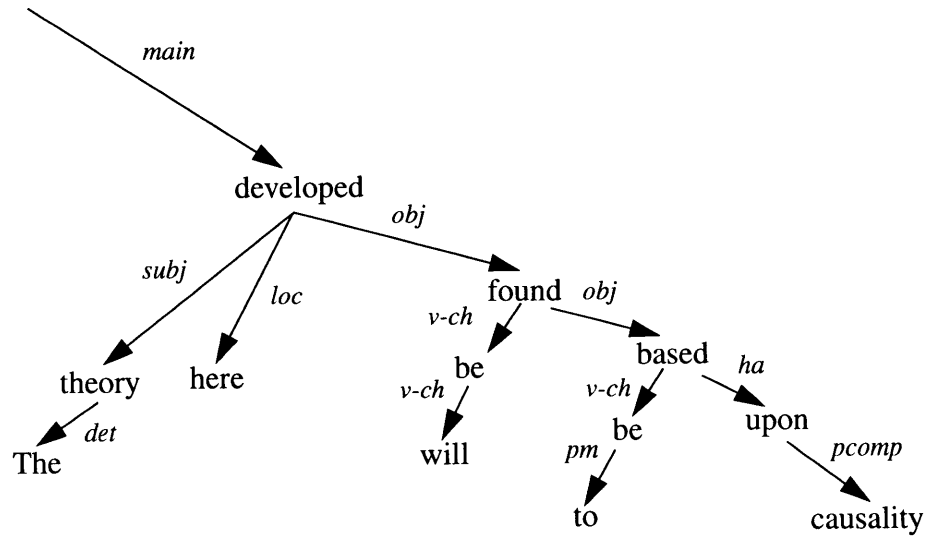


FIGURE 37. Dependency tree for the sentence “The theory here developed will be found to be based upon causality.”

The FDG system was used to parse the Globe and Clinton Chronica. Running on an

Tag	Explanation
main	main element (head verb for sentence)
v-ch	verb chain
pm	preposed marker
pcomp	prepositional complement
subj	subject
obj	object
comp	subject complement
oc	object complement
ha	heuristic high attachment (no rule applied)
det	determiner
mod	other postmodifier
cc	coordinating conjunction
attr	attributive nominal

TABLE 18. Principal link tagset for FDG (from Conexor oy, 1998b).

unloaded Pentium II (450 MHz) the parse took approximately two days per chronicon. Tapanainen and Järvinen (1997) have evaluated the FDG both for performance and precision of results. In their analysis, the system assigned functional dependencies to links with greater than 90% accuracy. For a newspaper corpus the system resolved subject dependencies with 95% precision, objects with 94%, and predcatives with 92% precision.

4.4.2 Noun phrase replicators

Noun phrases have been examined as lexico-syntactic replicators. Extracting noun phrases from the parse described above is a fairly straight-forward business. To

extract a noun phrase of maximum size I start with the head for each phrase which is tagged as a *main*, *obj*, *subj*, *comp*, or *pcomp*: in other words, the main verb of the sentence or of the subject, object, or complementing clause. Next, nodes in the dependency tree which are to the left of the head are examined for premodifiers. A premodifier will be marked as *attr* or potentially as *cc*. A coordinating conjunction (e.g., “and”, “or”) can string together words that are premodifying or postmodifying the head. After accumulating all premodifiers to the left of the head all links off and to the right are searched; these may be postmodifiers. Words tagged with *mod*, *pcomp*, and those marked as postmodifying coordinating conjunctions are accumulated. This algorithm results in noun phrases, such as the proper noun “Madeleine K. Albright.” But it also produces noun phrases, such as “realistic depictions of Native American women” (see Figure 38).

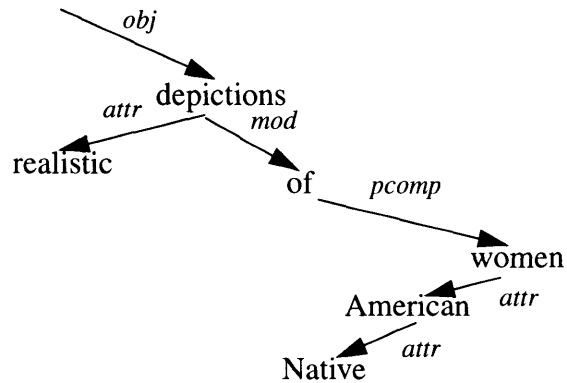


FIGURE 38. Dependency tree for “realistic depictions of Native American women.” The noun phrase algorithm will extract this entire string.

I recursively run the extraction algorithm on all sub-phrases extracting noun phrases of all sizes. Thus, for the example sentence above, I will extract: “realistic depictions of Native American women,” “depictions of Native American women,” “Native American women,” and finally, “American women.”

For the purpose of tracking noun phrase replicators prepositions and determiners are ignored and word ordering is normalized (lexicographically) so that the phrase of Figure 38 becomes “American depictions Native realistic women.” The intention is to accommodate variation in noun phrases that is not of strong semantic relevance. For instance, “big red car” and “red big car” would both be tracked as a single replicator. Though, similarly, “the cat on the mat” and “the mat on the cat”

Active Language Replicators

would be tracked as one replicator which arguably is not preferred. An inspection of the results, however, indicates that the phrase cohorts that are grouped as single replicators are mostly quite reasonable.

In Table 19 the top noun phrase replicators for the Globe and Clinton chronica are shown, after lemmatisation and word order normalization. The inconsistent state of capitalization is an outcome of the ENGTWOL lexical analysis. The system decapitalizes words that it does not think are proper nouns (e.g., “new” of New York) but

Globe chronicon	
Noun phrase	Count
ii war world	495
house official white	276
Albright Madeleine K. state	260
York city new	236
Monica lewinsky s.	215
Times York new	200
Clinton chronicon	
Noun phrase	Count
http://www.dejanews.com read search	971
deja discussion network news	966
http://clusterone.home.mindspring.com/ ignorance right wing	239
Linus Zimmerman f.	203
Dan dan.kimmel@worldnet.att.net	197
Harris John h.	194

TABLE 19. Top six lexico-syntactic noun phrase replicators (normalized form) from Globe and Clinton chronica. Clinton replicators are mostly from banner and footer text.

retains the original capitalization on words it thinks are proper names (e.g., “Madeleine” of Madeleine K. Albright). Why it retains the capitalization of “Albright” yet demotes “lewinsky” is a mystery to me. In any case, as long as this is performed consistently across all texts it will not influence the results.

For the Globe chronicon, recreating the actual noun phrases from the lemmatized and normalized expressions is trivial. We have: “World War II,” “White House official,” “State Madeleine K. Albright” which is a sub-phrase of “Secretary of State Madeleine K. Albright,” “New York City,” and so on.

The most frequent noun phrase replicators from the Clinton chronicon are not quite as satisfying as the Globe phrases. Most of them are due to confused parsings of banners and footers added to NetNews posts. For instance, the most frequently occurring two replicators are due to a text footer added to posts from the DejaNews system:

```
-----== Posted via Deja News, The Discussion Network ==-----  
http://www.dejanews.com/          Search, Read, Discuss, or Start Your Own
```

All of the top replicators seem to be the result of similar banner texts or are the email addresses or names of prolific authors, that do not lend themselves to sensible parsings.

Certainly, I could remove during pre-processing header and footer text such as these and other system produced tags, bylines, and so forth. However, as I demonstrate below, none of these strings are determined to have active evolutionary significance (but see Section 4.2.3 for a potential lexical counter example). In general, then, removing these system generated strings seems to be unnecessary.

4.4.3 Active noun phrase replicators

Let’s now turn to distilling active replicators from the two chronica. Table 20 shows the top noun phrase replicators from the Clinton/Lewinsky and Iraq clusters. Instead of presenting the normalized strings as was done in Table 19, the phrases have been restored to their attested form. For lexical replicators I considered lexemes that occurred at least 40 times and had correlation coefficients, $r \geq 0.5$, to be active. However, lexico-syntactic replicators occur far less often across a collection of texts and, on average, correlate less with the text rates (for that very reason if no other). So I will label lexico-syntactic replicators active if they occur at least seven

Active Language Replicators

times and have a correlation coefficient, $r \geq 0.4$. With that scheme in mind, I have

Clinton/Lewinsky cluster			
<i>r</i>	<i>p</i>	<i>n</i>	Noun phrase
0.5016	< 0.000001	6	White House intern
0.4792	< 0.000001	10	Lewinsky sexual relationship
0.4574	< 0.00001	8	his personal life
0.4373	< 000001	5	Dallas Morning News
0.4314	< 0.00001	162	Monica S. Lewinsky
0.4045	< 0.00001	9	Paula Jones lawsuit
0.4031	< 0.00001	58	former White House
Iraq cluster			
<i>r</i>	<i>p</i>	<i>n</i>	Noun phrase
0.5864	< 0.000001	5	US military strike
0.5522	< 0.000001	12	White House official
0.5211	< 0.000001	4	National Security Advisor Sandy Berger
0.5068	< 0.000001	4	Egyptian President Hosni Mubarak
0.4557	< 0.000001	25	Defense Secretary William S. Cohen
0.4451	< 0.000001	40	State Albright K. Madeleine
0.4391	< 0.00001	32	Iraqi President Saddam Hussein
0.4323	< 0.00001	4	Iraq military action possible

TABLE 20. Top active noun phrase replicators (attested form) from Clinton/Lewinsky and Iraq clusters.

histogrammed in Figure 39 the number of active lexico-syntactic replicators for the ten clusters of 200 texts or more within the Globe chronicon. It is interesting to compare this histogram with the one of Figure 30: clearly, the same three clusters have the most active set of replicators. Comparing Table 12 and Table 13 with Table 20 is also instructive: some active lexical replicators are apparently components of

the larger complexes which are resolved at the lexico-syntactic level. In some cases we may be simply revealing the same dynamic.

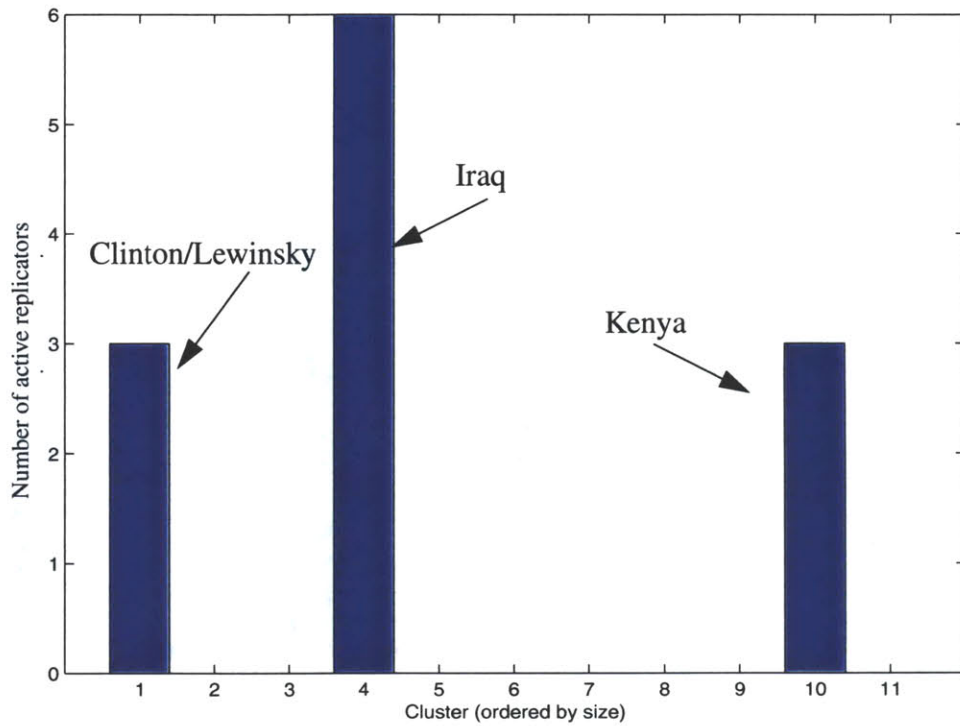


FIGURE 39. Number of active lexico-syntactic replicators for the ten clusters within the Globe chronicon with more than 200 texts.

The same analyses have been made with the Clinton chronicon. In Figure 40, I show the number of active replicators for each of the ten largest clusters. Interestingly, there is considerable variation between this graph and the lexical replicators of Figure 31. Table 21 shows the most active noun phrase replicators, in attested form, from the two most active clusters. Note that none of these active replicators are headers nor footers though many of the most common putative noun phrase rep-

licators were (Table 19). This supports my decision not to remove those texts as they have not entered into subsequent results.

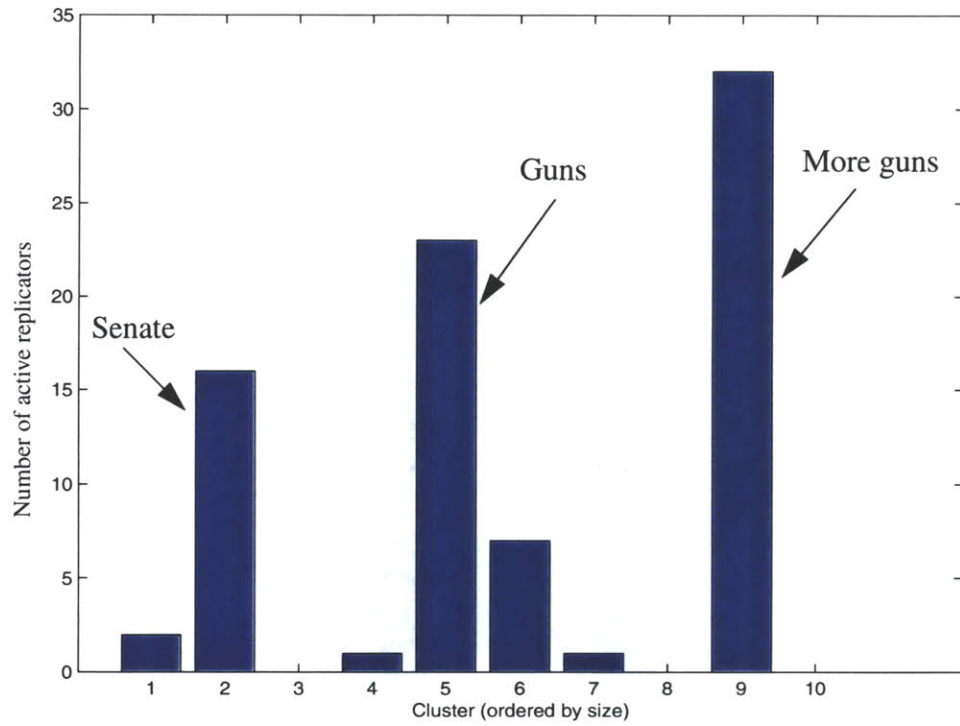


FIGURE 40. Number of active noun phrase lexico-syntactic replicators from the ten largest clusters within the Clinton chronicon.

Guns cluster			
<i>r</i>	<i>p</i>	<i>n</i>	Noun phrase
0.5322	< 0.000001	14	concerns about Federal interference
0.5021	< 0.000001	9	U.S. v. Miller
0.5021	< 0.000001	8	Federal court decision holding
0.5021	< 0.000001	9	inhibition of Second Amendment
More guns cluster			
<i>r</i>	<i>p</i>	<i>n</i>	Noun phrase
0.0668	< 0.000001	8	Ford office building
0.6208	< 0.000001	106	Gun Control Part Three
0.5180	< 0.000001	35	legal justifiable homicide
0.5070	< 0.000001	12	potential bodily harm
0.5025	< 0.000001	9	immediate threat to your life

TABLE 21. Top active noun phrase replicators (attested form) from two gun control clusters within Clinton chronicon.

4.4.4 Summary and evaluation

Not surprisingly, large noun phrases replicate less frequently than lexical elements. This alone may account for the generally lower correlations between the noun phrase and the lexical replicators. In many cases, the lexical and lexico-syntactic method describe what seem to be the same or similar units of replication. For instance, within the Clinton/Lewinsky cluster the word “Lewinsky” reoccurs 309 times and has a correlation of $r = 0.5270$. In the same set of texts, the noun phrase “Monica S. Lewinsky” occurs 162 times with $r = 0.4314$. Thus, the 162 occurrences of the noun phrase account for half of the occurrences of “Lewinsky” on its own. Does the increase in use of only the last name account for all of the increase in autocatalysis (in Section 5.8 I discuss the appropriate size of units of analysis — words, phrases, etc.)?

A comparison of the noun phrase replicators with lexical co-occurrence shows, I believe, that noun phrases are more coherent and useful units of analysis. These replicators enjoy similar levels of active correlation, in general, but are more clearly distilled from the chronica and can be read directly and understood.

But, why do certain noun phrases replicate with such strong autocatalytic correlation whereas minor variants to it fail to do so? For instance, “White House official” is an extremely successful replicator from the Clinton/Lewinsky cluster; but “White House spokesman,” which occurs 230 times throughout the chronicon, does not actively replicate. In these texts “official” is often used to describe an anonymous source, whereas “spokesman” is always an official source on the record. This cultural difference may be the source of their variance in autocorrelation — officials may be better and more evocative sources than spokesmen.

4.4.5 SVO replicators

Other lexico-syntactic replicators were investigated. In particular, I explored subject/verb/object (SVO) trigrams. Since these three components make up the foundation of the English sentence I hoped that they would perform well as replicators.

The SVO extraction algorithm is fairly straightforward. From the dependency tree a link labeled *main* points to the main verb; a subject usually sits to the left and an object to the right — picking off these components is simple. In Figure 37, the main verb is “developed,” the subject is “theory,” and the object is “found.” The SVO trigram for this sentence is “developed/theory/found.” However, the main verb often is not the only verb of a sentence. A link marked *subj*, *obj*, *pcomp*, or *comp* may point to the verb of a phrase that in turn may have a subject to the left and an object to the right. Consider the dependency tree for the sentence “I saw the bear eat the man,” as shown in Figure 41. The object of the sentence is a phrase with an embed-

ded SVO trigram, namely, “bear/eat/man.” My algorithm will extract this SVO trigram along with “I/saw/eat.”

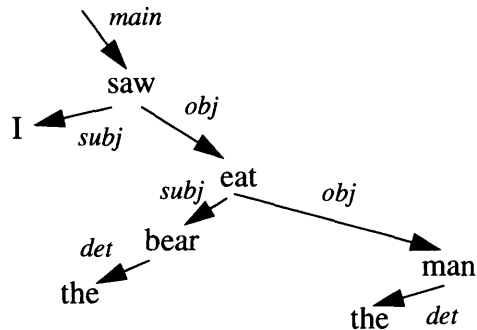


FIGURE 41. Dependency tree for “I saw the bear eat the man.”

Table 22 depicts the top SVO replicators from the Globe and Clinton chronica. Clearly, lots of folks are thinking and saying things but not much information can be gained from these replicators. The personal pronouns and non-specific verbs produce highly general SVO trigrams. “Clinton/say/be” provides one counterexample where at least we are given a specific subject. The “search/start/own” trigram is an outcome of the same DejaNews banner that was described in Section 4.4.2.

I first explored SVO replicators within the Globe chronicon. These lexico-syntactic features did not produce favorable results. Table 23, shows the *only* three active SVO replicators, those with an $r \geq 0.4$ and $n \geq 7$, from the entire Clinton chronicon. The clear conclusion is that SVO trigrams, as currently implemented, are not very meaningful active replicators within the Globe chronicon. There are very few strongly correlated SVO trigrams, and the few that are identified are not meaningful; they are composed primarily of personal pronouns and the verb “say” or

Active Language Replicators

“think.” (“Ginsburg,” the only proper noun in the lot, was the lawyer for Monica Lewinsky.)

Globe chronicon	
SVO trigram	Count
he say be	1900
I think be	818
she say be	534
official say be	486
he say have	347
Clinton say be	204
Clinton chronicon	
SVO trigram	Count
I think be	1032
search start own	955
you think be	295
I know be	273
he do it	244
I believe be	238

TABLE 22. Top six lexico-syntactic SVO replicators (normalized form) from Globe and Clinton chronica.

Cluster	<i>r</i>	<i>p</i>	<i>n</i>	SVO trigram
Clinton/Lewinsky	0.4112	< 0.000001	8	Ginsburg say be
Ireland/UK	0.4355	< 0.000001	19	I think be
Kenya	0.5509	< 0.000001	21	he say be

TABLE 23. The only three active SVO trigram replicators from Globe chronicon. Not very compelling results.

The SVO replicators from the Clinton chronicon appear to be far more meaningful. Figure 42 shows a bar graph for the ten largest clusters; there are a considerable number of active SVO replicators within this chronicon. In Table 24 the top one or two active SVO replicators from each of the five clusters that had any is shown. Although some of these replicators are fairly general (e.g., “you/defend/he”), they still carry more semantic weight than those from the Globe chronicon (“you/defend/he” came from sentences where people were being criticized for defending Bill Clinton). Some are quite specific, such as “Clintonphobe/grind/tooth,” which

comes from the sentence “Cold Bastard Bill Makes Clintonphobes Grind Their Teeth AGAIN!”

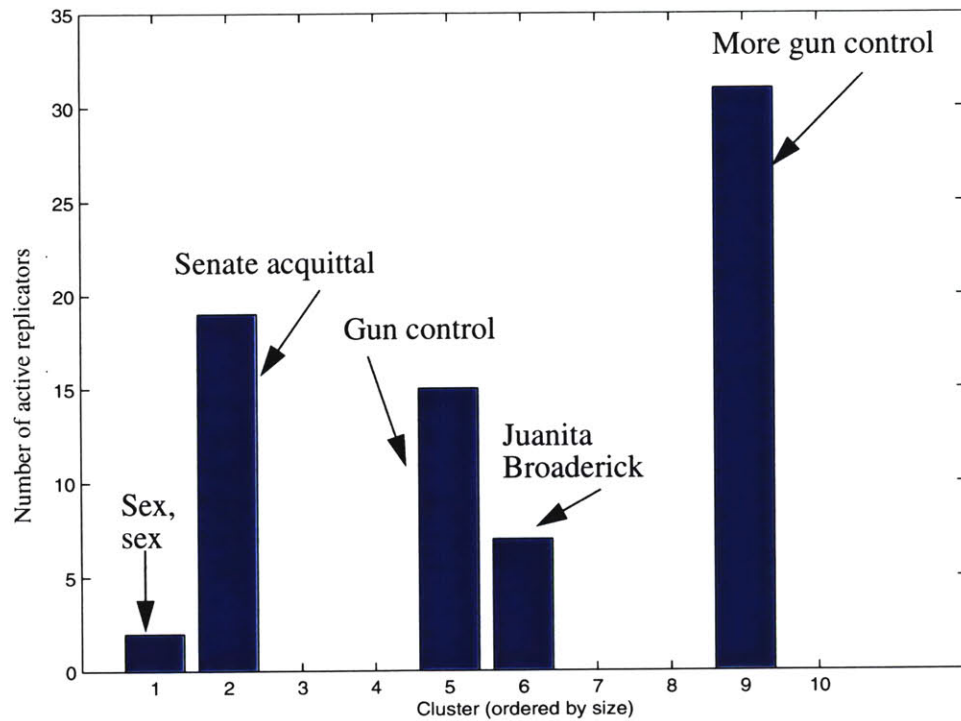


FIGURE 42. Number of active SVO lexico-syntactic replicators from the ten largest clusters within the Clinton chronicon. Results are much more compelling than for the Globe chronicon.

I believe the SVO trigrams within the Clinton collection, compared to the Globe collection, are more meaningful, in large part, due to the frequent verbatim copying

across in-reply-to threads. And some of these reoccurring sentences have auto-catalytic significance and generate active SVO trigrams.

Cluster	<i>r</i>	<i>p</i>	<i>n</i>	SVO trigram
Sex&sex	0.4700	< 0.000001	14	Clintonphobe grind tooth
Sex&sex	0.4121	< 0.000001	8	it fuck wad
Senate	0.5671	< 0.000001	14	we say limit
Guns	0.5489	< 0.000001	8	debate have do
Guns	0.5191	< 0.000001	11	congress make law
Broaderick	0.4691	< 0.000001	10	you defend he
Broaderick	0.4581	< 0.000001	27	interview see woman
More guns	0.7068	< 0.000001	9	you get dick

TABLE 24. Top active SVO trigram replicators from five clusters of the Clinton chronicon.

For the Globe chronicon, I believe the SVO trigrams have not proven effective because the highly recurrent ones are too general to carry any semantic weight. Thus, “he/say/be” occurs frequently as an SVO trigram across a wide range of topics within the chronicon. Some specific cases of this replicator may be active and others not, and when they are combined into a single unit of analysis the end result is a trait that does not do well, overall, in correlating with the volume of texts. The two cases in the Globe chronicon that were reasonably active, see Table 23, were cases where “he” or “I” referred to a particular individual enough of the time to act as an autocatalytic trait.

The obvious fix to this problem is, first, to resolve the personal pronouns. I have begun to explore anaphoric resolution algorithms (Hobbs, 1978; Asher & Wada, 1988; Lappin & Leass, 1994; B. Baldwin, 1995; Rocha, 1997); these would insert the referent in place of all of the personal pronouns and thus narrow the replicators to their specific targets, which should sort out the active replicators from the non-correlated ones. It also would be useful to determine more semantically relevant verbs (instead of “think,” “be,” or “say” for instance). This remains an open research question.

In summary, while SVO trigrams may make up reasonable lexico-syntactic replicators, my naive extraction algorithm produces strings that are far too general, at least within the Globe chronicon, to demonstrate regular and meaningful active replication.

4.5 Syntax

The final level of analysis considered is syntactic. I have examined the replication of structural patterns through the chronica. Without doubt, syntactic patterns are reoccurring with frequency throughout any collection of texts; but are some of these patterns actively replicating?

The structural patterns I examined are part of speech (POS) n -grams. Each word in the collections was tagged with its part of speech (noun, verb, etc.) and then grouped, by tag, into strings of size $n = 4$. The EngCG-2 constraint-based morphological tagger (Samuelsson & Voutilainen, 1997), which is based on the original EngCG tagger (Karlsson, Voutilainen, Heikkillä & Antilla, 1995), was used to tag words with their part of speech. (In Chapter 3, I referred to the lexical analyzer ENGTWOL which I have used to lemmatize the chronica; this is a sub-process of the EngCG-2 tagger.) The tagging system first tokenises the text by identifying proper word breaks, grouping common collocations, and so forth. It then performs a morphological analysis making use of a large lexical database of over 56,000 words (Conexor oy, 1998a). For words that are not in the database, it employs a rule-based heuristic system (Voutilainen, 1995a). (This is the lemmatisation procedure described in Chapter 3.) At this point, each word has been assigned a lemma. Each lemma in the lexicon is coded for the parts of speech it can appear in. Often, words can appear in a variety of places within a sentence, thus the lexical entries are ambiguously coded. Moreover, words unknown to the database or used in unusual circumstances might not be coded at all, or may be coded inconsistently. The final step for EngCG-2 is to resolve any ambiguities and arrive at a final part-of-speech coding for each token (Voutilainen, 1995b). To do this, a grammar of over 4,000 rules is employed; these rules discard illegitimate tags, based on local and global context conditions. EngCG-2 is reported to be highly accurate. In tests, the system fully disambiguates 96% - 98% of all input words, leaving only a small percent of the words with multiple tags. The error rate, those words assigned erroneous tags, is 0.2% - 0.4% on test input (Conexor oy, 1998a).

Part of speech	Explanation
N	noun
ABBR	abbreviation
A	adjective
NUM	numeral
PRON	pronoun
DET	determiner
ADV	adverb
ING	ING-form
EN	EN-form
V	verb, finite or infinitive
INTERJ	interjection
CC	coordinating conjunction
CS	subordinating conjunction
PREP	preposition
NEG-PART	“not” or “n’t”
INFMARK>	“to,” “in+order+to,” etc.

TABLE 25. EngCG-2 part of speech tagset (from Conexor oy, 1998a).

Subfeature	Explanation
NOM	nominative
GEN	genitive
ACC	accusative
SG, PL, SG/PL	singular, plural, singular or plural
SG1, SG2, SG3	singular first, second, third person
PL1, PL2, PL3	plural first, second, third person
ABS	absolute
CMP	comparative
SUP	superlative
DEM	demonstrative
FEM	feminine
MASC	masculine
INF	infinitive
IMP	imperative
PRES	present tense
SUBJUNCTIVE	subjunctive
PAST	past tense
AUXMOD	modal auxiliary

TABLE 26. EngCG-2 selected subfeature tags (from Conexor oy, 1998a).

Table 25 shows the part of speech tagset employed by EngCG-2, and Table 26 shows the subfeature set. Words are tagged with the fullest set of subfeatures appropriate and available from the lexicon and constraint grammar. I have tagged the example sentence from Figure 37 and show the output of EngCG-2 in Figure 43. The first field shows the words as they appear within the actual text (quoted and enclosed by angle brackets). The second field displays the lemma for the word,

enclosed by quotation marks, followed by the part of speech tag and any subfeatures that are identified for this word. For instance, the word “The” is assigned the lemma “the” and tagged as a determiner of singular or plural number. Note the word “developed,” is lemmatized as “develop.” The part of speech tagging is ambiguous, marking the word as either an EN-form or as a past tense verb. The EN tag refers to the past participle and gains its name from the common word ending for the past participle in English, such as the “en” in “taken.” In this sentence, the correct tag for “developed” is EN.

"<The>"	"the" DET SG/PL
"<theory>"	"theory" N NOM SG
"<here>"	"here" ADV
"<developed>"	"develop" EN "develop" V PAST
"<will>"	"will" V AUXMOD
"<be>"	"be" V INF
"<based>"	"base" EN
"<upon>"	"upon" PREP
"<causality>"	"causality" N NOM SG
"<\$.>"	

FIGURE 43. Part of speech tags from the EngCG-2 analyzer for the sentence “The theory here developed will be based upon causality.”

After tagging is complete, the algorithm to extract POS *n*-grams is simple; the only complication that might occur is due to ambiguity and the use of subfeatures. When assembling the 4-grams all possible strings of tags are considered and I recursively

group together each tag for those assigned ambiguously. In the example in Figure 43, I would consider both the feature “DET, N, ADV, EN” and “DET, N, ADV, V” for the two ambiguous coding of the token “developed.” (Note that the recursive element of this procedure occurs when there are multiple ambiguities within a single four word window.) I have discarded the subfeatures here. I do consider replicators composed solely of the part-of-speech tag without any of the subfeatures. But I also employ the subfeatures, so I would in addition code this sentence with the string, “DET SG/PL, N NOM SG, ADV, EN.” This extraction algorithm proceeds along the texts extracting all part-of-speech replicators for all contiguous windows of length four.

In Table 27 the top POS *n*-gram replicators for the Globe and Clinton chronica may be seen. The top syntactic replicator for both collections is “N, PREP, DET, N.” This would correspond to a phrase such as “cat in the hat.” Note that none of the top replicators contain subfeatures. This is not surprising as the subfeatures only narrow or reduce the occasions that a tagging might occur.

This analysis immediately suggests a potential stylostistical measure for collections of texts. The more diversity in POS *n*-grams, the more structural variety is present within a collection. If most syntactic replicators are covered by a small number of POS *n*-grams, then less structural diversity exists within the texts than if a large number of *n*-grams appear, on average, less frequently. For the Globe chronicon, the average POS replicator appears 26.92 times ($\sigma = 413.55$), for the Clinton chronicon, the average is 21.53 ($\sigma = 230.06$). These numbers are clearly sensitive to the relative size of each collection — more text will mean that the *n*-grams are expected to occur, on average, more often. In Figure 44, I histogram on a log-log plot the relative number of times each POS *n*-gram appears in the two chronica. I know exactly how many times each syntactic replicator has appeared in each chronicon. Dividing this number by the total number of POS *n*-gram’s across the entire collection, I am left with the proportion of text which makes use of each syntactic structure. I have histogrammed these proportional values into buckets for those that occur often and for those that are rare. A large number of POS *n*-grams (nearly 10^6) occur between one and ten times throughout the collection (for instance <ADV-N> N NOM SG, NUM CARD, PRON DEM SG, PRON PERS FEM NOM SG3 occurs once in the Globe chronicon). These cases are plotted on the left of the graph. Similarly, a very small number of POS *n*-grams occur hundreds or thousands of times; this is the right of the graph.

Globe chronicon	
POS <i>n</i> -gram	Count
N, PREP, DET, N	104625
N, N, N, N	91236
N, PREP, N, N	58942
PREP, DET, N, N	58924
PREP, DET, A, N	55090
PREP, DET, N, PREP	54285
Clinton chronicon	
POS <i>n</i> -gram	Count
N, PREP, DET, N	48930
V, DET, N, PREP	28078
PREP, DET, N, PREP	27971
DET, N, PREP, DET	26692
DET, N, PREP, N	26096
PREP, DET, A, N	25094

TABLE 27. Top six POS *n*-gram syntactic replicators from Globe and Clinton chronica.

Given that this is a log-log graph even fairly minor variations express meaningful differences. The dotted line is the plotted histogram for the Clinton chronicon. That it is slightly elevated from the graph for the Globe chronicon demonstrates that there is greater diversity in POS *n*-grams for that collection, because more POS *n*-grams, proportionally, show up a very few number of times.

It has been demonstrated that POS trigrams are effective means to discriminate between authors (Milic, 1966; Oakes, 1998). In one example, Milic attempted to determine if Jonathan Swift (most famous for his *Gulliver's Travels*) was the author

of *A Letter of Advice to a Young Poet*, whose authorship at the time was in dispute. He compared this text to known writings of Swift, as well as to writings attributed to other contemporary authors, and found, first, that the most common POS trigram was PREP, DET, N, which is consistent with my finding for 4-grams. Second, he found that the total number of POS trigrams was an effective metric to distinguish authors.

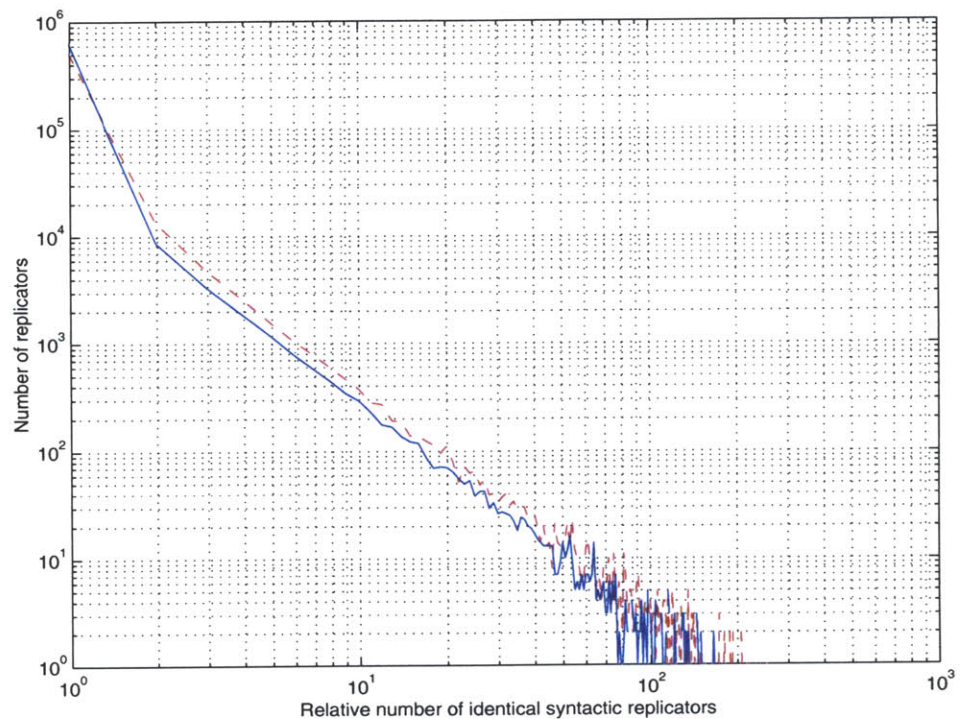


FIGURE 44. Log-log graph of histogram for relative number of times each POS n -grams occurs within the Globe (solid) and Clinton (dashed) chronica.

In Figure 45, I show the histogram of correlation coefficients for POS n -gram replicators from the Clinton/Lewinsky cluster of the Globe chronicon. The mean correlation coefficient is 0.0725 ($\sigma = 0.1289$). This distribution does not look considerably different than that for lexical replicators shown in Figure 46 of Chapter 3. However, because there are far more POS n -gram replicators distilled from the chronoroom compared even to lexical replicators (let alone lexico-syntactic), I

must be extra cautious to protect against chance and artifact (if for no other reason than chance events are more likely to occur when there are more events).

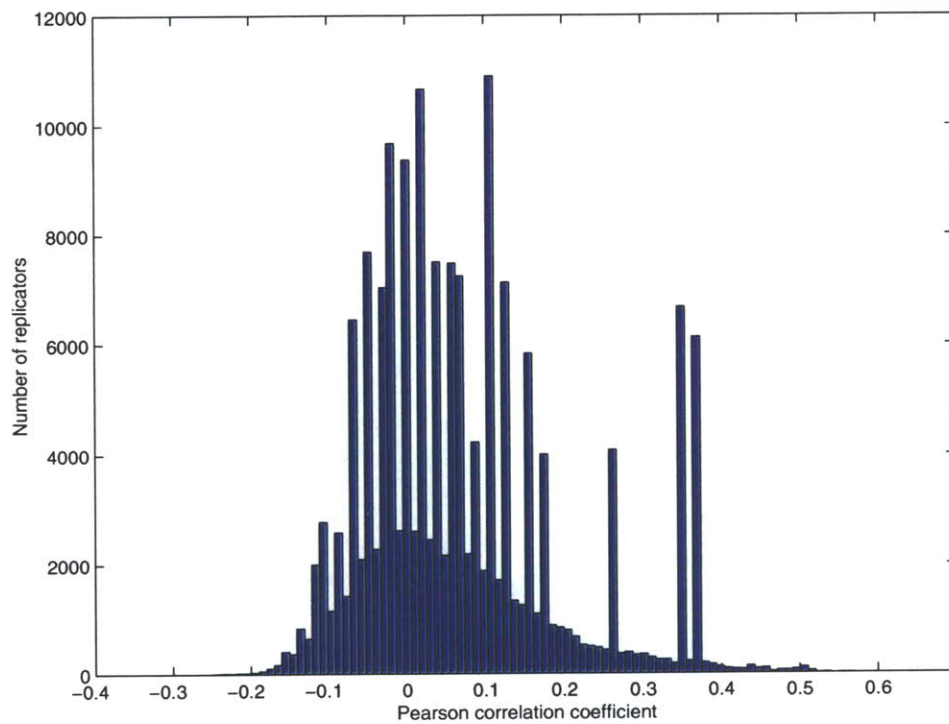


FIGURE 45. Histogram of correlation coefficients for all POS n -gram replicators from Clinton/Lewinsky cluster of Globe chronicon.

Only those syntactic replicators that occur at least 100 times are shown in Figure 46; this is a reasonable number of reoccurrences to require, given that we still have many replicators that are this frequent. This distribution looks quite normal ($\bar{x} = 0.0163$, $\sigma = 0.1195$). Indeed, the heavy Gaussian quality of this graph gives me

pause. I worry that those replicators with high r (there are four replicators with $r \geq 0.4$) might simply be tails to a normal distribution.

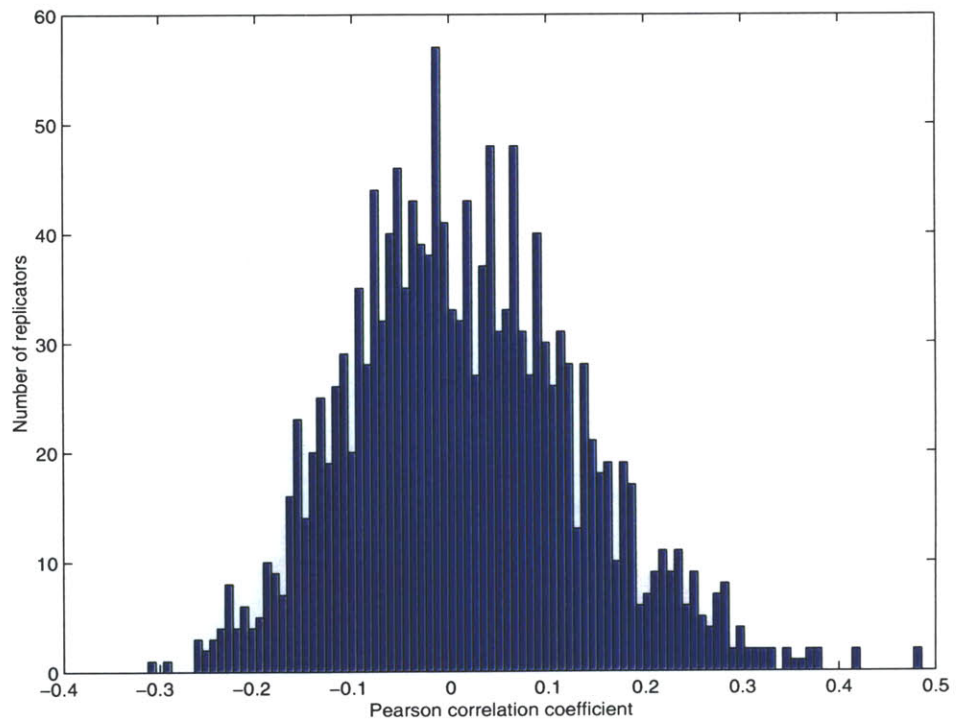


FIGURE 46. Histogram of correlation coefficients for POS n-gram replicators that occur at least 100 times in Clinton/Lewinsky cluster of Globe chronicon.

The larger number of POS replicators compared to the other linguistic features will, with high probability, increase the number of statistical outliers, suggesting that

maximum caution should be applied in analyzing the results (see Table 28).

Chronicon	Replicator type	Total number of replicators
Globe	lexical	44,612
	lexico-syntactic (noun phrase)	415,688
	lexico-syntactic (SVO trigram)	234,355
	syntactic (POS 4-gram)	656,434
Clinton	lexical	28,185
	lexico-syntactic (noun phrase)	85,147
	lexico-syntactic (SVO trigram)	64,919
	syntactic (POS 4-gram)	546,745

TABLE 28. Total number of lexical, lexico-syntactic, and syntactic replicators from Globe and Clinton chronica. Note the relatively large number of POS 4-grams.

Table 29 shows the four syntactic replicators with $n \geq 100$ and $r \geq 0.4$ from the Clinton/Lewinsky cluster. These counted as active under the lexical regime. For every other cluster with over 100 texts within the Globe and Clinton chronica there were only two other replicators that met these requirements of $n \geq 100$, $r \geq 0.4$; they too are shown in Table 29.

Globe chronicon				
Cluster	<i>r</i>	<i>p</i>	<i>n</i>	POS 4-gram
Clinton/Lewinsky	0.4855	< 0.00001	224	V, PRON, PRON, V
Clinton/Lewinsky	0.4823	< 0.00001	138	PRON, V, ADV, ADV
Clinton/Lewinsky	0.4216	< 0.0001	143	PRON, V, EN, PRON
Clinton/Lewinsky	0.4155	< 0.0001	209	PRON, V, EN, DET
Iraq	0.4359	< 0.0001	154	N NOM SG, N NOM SG, ABBR NOM SG, N NOM SG
Clinton chronicon				
Cluster	<i>r</i>	<i>p</i>	<i>n</i>	POS 4-gram
Broaderick	0.4666	< 0.001	139	V, DET, N, CC

TABLE 29. Only POS *n*-gram replicators from all large clusters of the Globe and Clinton chronica with $n \geq 100$, $r \geq 0.4$.

4.5.1 Evaluation

Consider the four most correlated replicators from the Clinton/Lewinsky cluster. They all contain the sub-feature of PRON, V; perhaps these features truly are auto-catalytic, due to some quality they possess (in contrast to their simply being in the tail of a distribution).

To remove any doubt that statistical artifact is at play, Figure 47 shows the timeseries for both the volume of texts, over time, within the Clinton/Lewinsky cluster and the relative presence of the V, PRON, PRON, V *n*-gram feature. The strong correlation between these two timeseries is immediately obvious. Note that

these two series appear to be reasonably stable; they do not exhibit significant trend nor drift.

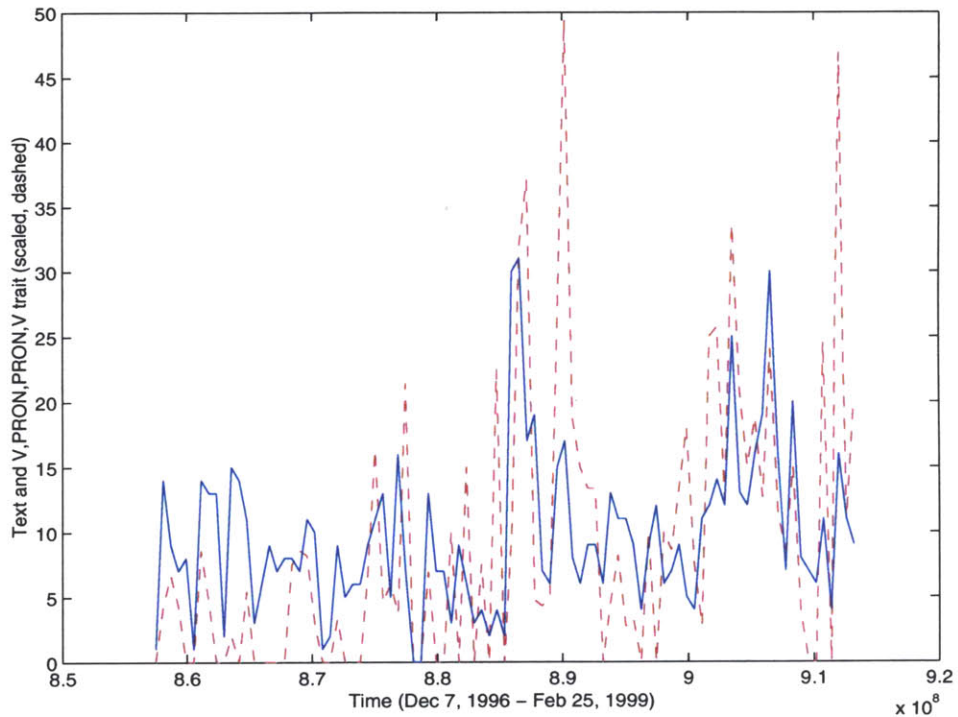


FIGURE 47. Syntactic replicator V, PRON, PRON, V and volume of texts published within the Clinton/Lewinsky cluster of the Globe chronicon ($r = 0.4855$).

I mentioned in Section 3.11.2 that autocorrelation was another worry when attempting to correlate timeseries. An investigation of the spectral components is the single most powerful tool we have to detect such autocorrelations (McCleary & Hay, 1980; Wei, 1990). The real portion of the Fast Fourier Transform for both the text volume and trait timeseries appears in Figure 48. Both of these spectral graphs

look good; they are nicely centered about the zero origin and do not have significant spikes or jumps in them. Thus these timeseries do not require pre-whitening.

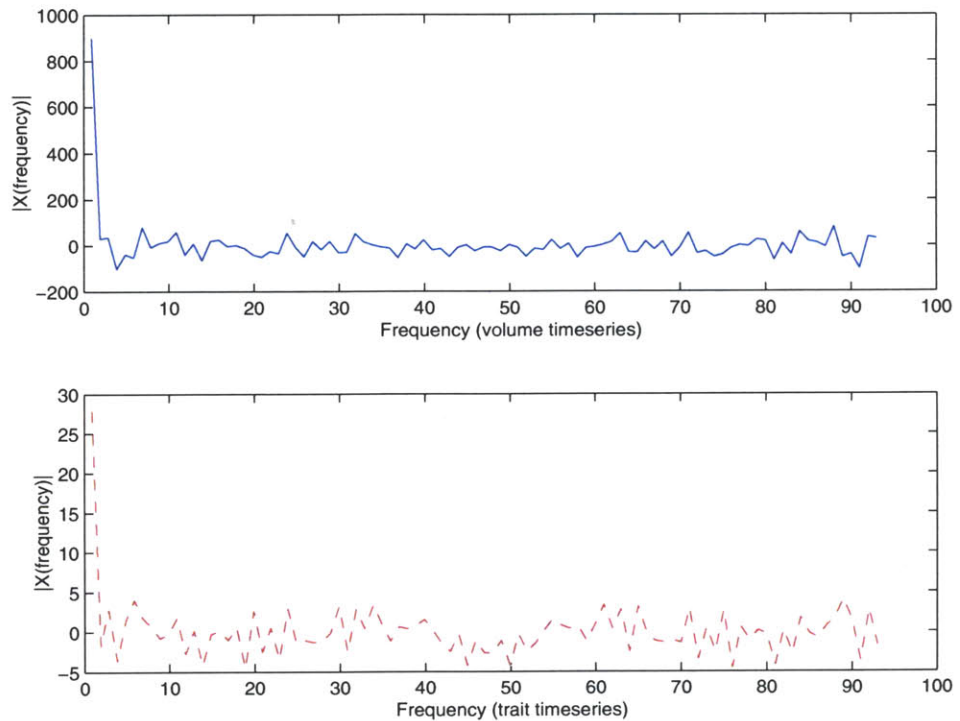


FIGURE 48. Fast Fourier Transform of the two timeseries of Figure 47. The spectral components are free from spikes and symmetric about the middle.

The above analysis suggests that the strong correlation of the V, PRON, PRON, V replicator is not a statistical artifact, but it can still be due to statistical chance. For the other linguistic levels, a qualitative analysis was the only real means to protect against statistical chance, and we should look for the same qualitative arguments here. Are there functional or adaptive explanations for the phenomena exhibited? (This general theme will be taken up in Section 5.6.) Qualitative explanations are difficult to assess at the syntactic level, since it is hard to evaluate the functional role they may play in some piece of language: what problem is solved with V, PRON, PRON, V. Why would the feature V, PRON, PRON, V be autocatalytic (or better perhaps, why is the tuple PRON, V autocatalytic)?

One possible explanation (beyond that of chance) is that all of these statistical features are simply acting as traces for the real traits which are occurring at other linguistic levels, namely lexico-syntactic. Perhaps these syntactic features all correspond to the same (more or less) set of words, and it is at the level of these words that we have our real explanation and autocatalysis. Table 30 shows the lemmatized words and punctuation associated with the V, PRON, PRON, V replicator for the first eight occurrences in the Globe chronicon. Clearly, each occurrence of this replicator is not due to the same set of words reappearing.

V, PRON, PRON, V POS 4-grams
say what they should
take anything they say
get it. I be
say. "What we should
do what he think
do what he's ("s" is lemmatized as "be" or "have")
suggest something that be
be something he can

TABLE 30. The lemmatized words associated with the first eight occurrences of the V, PRON, PRON, V syntactic replicator.

Considering this evidence, my final conclusion is: "definite maybe." I do think it is possible that the V, PRON, PRON, V syntactic replicator (or PRON, V subfeature), due to some quality or design feature it possess, is active within the Globe chronicon. I have evidence that the correlations for these features within the Clinton/Lewinsky texts are not due to artifact. And I do believe it is possible that it is not due to chance. But currently I do not have a qualitative argument as to how or why these replicators would be active.

4.6 Summary

The Microevolutionary Language Theory postulates that complex design accumulates at the simplest levels within human natural language due to the evolutionary algorithm. In the previous chapter I described the CAMEL software system. In this

chapter I reported on active replicators at a number of simple levels within natural language. I distilled autocatalytic replicators at the lexical, lexical co-occurrence, lexico-syntactic, and perhaps at the syntactic level. These hundreds of examples of autocatalytic replicators offer support to the Microevolutionary Language Theory (developing this argument will be a central feature of the next chapter).

These active replicators are generally the most central, provocative, controversial, and evocative elements in the texts. In reporting on the Clinton/Lewinsky scandal, for instance, the active replicators are about the central themes: morality, denials and the truth, sex and an affair, and the exposure of Clinton's personal life. The most provocative and central players in this scandal also come out as active replicators: Clinton, Lewinsky, Ginsburg. And we also see considerable dysphemism: dick, cunt, and fuckwad, as active replicators.

The CAMEL system set out to support the Microevolutionary Language Theory by distilling evolutionarily significant autocatalytic replicators. But a read through of these results makes it clear that identifying active replicators is tantamount to distilling the "hot" and provocative topics, terms, and people within a chronicon — those topics, terms, and people which generate a strong response from the socio-cultural environment.

The Microevolutionary Language Theory states that complex functional design accumulate at the simplest level within language due to the evolutionary algorithm. This claim rises from a novel conceptual integration of contemporary evolutionary theory with corpuslinguistic models of language use.

Central to contemporary evolutionary theory is the replicator; this model now stands as the conventional orthodoxy within the evolutionary community (Plotkin, 1994; Pinker, 1997). In the previous two chapters I described a system to distill active replicators from within text collections. But I have yet to firmly set those results into the framework of contemporary evolutionary theory and to describe exactly how it may lead to an understanding of complex functional design. It is in this chapter that I turn to these points.

I will begin with an overview of the two principal systems under which contemporary evolutionary theory is organized: the Lewontin-Campbell computational theory and the Dawkins-Hull typological theory. I will place the active language replicator model into these frameworks and, in so doing, will argue for a corollary to Campbell's Rule which states that cultural evolution is the same as organic evolution, that is, they both are running the same algorithm. This result, which I call the Microevolutionary Language Corollary, states that language evolution runs the same algorithm as organic evolution.

This chapter then turns to a variety of important theoretical issues. First, I look into units of selection within language and consider what is the central beneficiary of the evolutionary process. The replicator-eyed viewpoint offers what may be a surprising result. Next I consider the link between microevolution and the accumulation of complex adaptive design; here I finally show how the active replicator model of language offers an explanation for design.

In the final three sections I touch briefly on some controversial areas of current evolutionary theory. First, I examine selection, arguing that it is a powerful force within human natural language. I then consider what is the exact size and level of active language replicators: words, collocations, phrases, etc. And, finally, I argue for a link between the microevolutionary language replicator model and the macroevolutionary patterns studied within traditional historical linguistics.

5.1 Campbell's Rule

The psychologist Donald T. Campbell established a programme aimed at understanding the evolution of culture and knowledge, a research endeavor he referred to as "evolutionary epistemology" (D.T. Campbell, 1974). His work is currently the most influential in this area, though it follows a research direction that goes back to (and even predates) William James (1880), James Mark Baldwin (1896/1996), Karl Popper (1972) who applied these views in particular to science, and Jean Piaget (1980) who argued (without success) for a radical "genetic epistemology" (see Plotkin, 1994).

Campbell spoke persuasively to many issues of an evolutionary culture theory. Not least amongst them was the observation that cultural evolution and organic evolution are both examples of a single, larger, over-arching process. Durham (1991, p. 187) has proposed calling this "Campbell's Rule."

Consider the diagram of Figure 49. Each link is labeled with the relation, or lexical function (Mel'cuk, 1988), that associates the connected words. Thus, if *a dog is a type of animal*, then *dog* is an hyponym to *animal*, and *animal* is an hypernym to *dog*. Further, if *a cat is also an animal*, then *cat* and *dog* are in a particular relation to each other. This relation has on occasion been called "co-hyponymy" (Matthews,

1997); but I instead propose *isonym*, from the Greek root for equal or parallel, as a more appropriate and evocative term for this relationship.

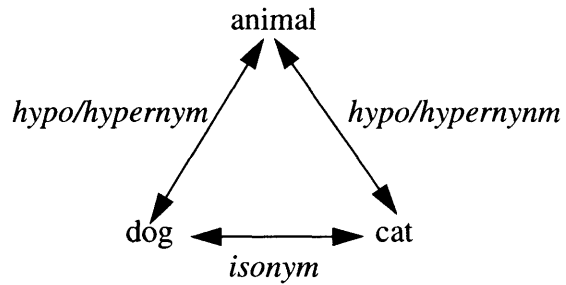


FIGURE 49. Relation, or lexical function, between “dog,” “cat,” and “animal.”

Campbell wrote that “the analogy to cultural accumulations [is not] from organic evolution per se; but rather from a general mode... for which organic evolution is but one instance” (1965, p. 26). These relations I’ve pictured in Figure 50. With this bit of background, Campbell’s Rule can be simply stated:

Campbell’s Rule: Cultural evolution and organic evolution are isonyms to each other (and hyponyms to a general process of evolution).

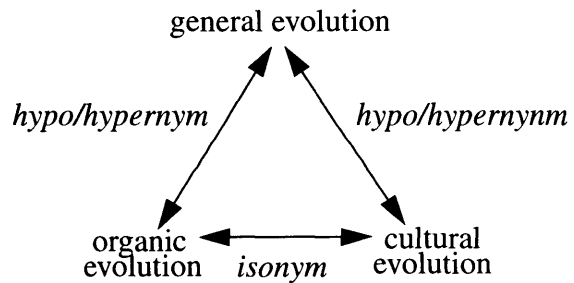


FIGURE 50. Campbell’s Rule states that organic evolution and cultural evolution are isonyms to each other.

My goal now is this: to argue from the results of the previous chapter that language evolution is also an isonym to organic evolution. In other words, I intend to demonstrate a Microevolutionary Language Corollary to Campbell's Rule. In order to accomplish this, however, I must first describe what is meant by general evolution. In so doing I will place the active language replicator model into the framework of the Lewontin-Campbell computational theory of evolution.

5.2 The Lewontin-Campbell Computational Theory

A *computational theory* of some natural phenomena is an algorithmic explanation of that phenomena — what information and representations are required, what computations are involved (Marr, 1982; Richards, 1988; Pinker, 1997). Recently, a number of influential computational theories of evolution have been put forth. While they have almost always been proposed as general theories of evolution, they are offered within the context of some specific substrate of evolution, namely, molecular (Eigen, 1992), cultural and psychological (D.T. Campbell, 1960), organic (Lewontin, 1970), and artificial (Holland, 1992).

These four computational theories differ in ways that are not particularly substantive. And certainly each posits a general process of evolution: an over-arching theory that Richard Dawkins (1983) has aptly named Universal Darwinism. Plotkin (1994) has taken to calling this general computational theory of Universal Darwinism the *Lewontin-Campbell approach* (he does not, though, observe that this is an algorithmic explanation). I do like the idea of singling out, in particular, these two researchers, and so am happy to meet Plotkin half-way and refer to it as the "Lewontin-Campbell computational theory."

What exactly is the Lewontin-Campbell computational theory of evolution? Campbell described the algorithm as a process of blind-variation-selective-retention, Lewontin pitched it as variation, reproduction, and heritability, and Plotkin re-described it as generate-test-regenerate (g-t-r). What is clear, is that the algorithm requires:

1. a population of individuals over time with a correlation in heritable traits,
2. expressed variation,
3. and a covariance between variants and the success in time of the population (Pocklington & Best, 1997).

These are the necessary and sufficient conditions for the evolutionary algorithm. If they are met, evolution will follow (Fisher, 1912).

There can be no doubt that this is a computational theory in the sense I have in mind — an algorithmic explanation of some natural phenomena. The simplest way to establish this is to note that it describes quite nicely the basic Genetic Algorithm (GA) (e.g., Goldberg, 1989; Holland, 1992; Zbigniew, 1992; Back, 1996; Mitchell, 1996). The GA is a well known computer algorithm used primarily on problems of optimization and search. Holland (1992) has shown (through reliance on the multi-armed-bandit problem) that the GA, and therefore the Lewontin-Campbell evolutionary algorithm as well, optimally trades-off the two goals of search within a non-linear or epistatic landscape: the *exploration* of new potential answers and the *exploitation* of existing known solutions. This amounts to a *post-hoc* mathematical demonstration that nature optimally searches its problem space!

5.3 Microevolutionary Language Corollary

Can I identify the three necessary and sufficient conditions of the Lewontin-Campbell algorithm within my active language replicator model? If I can demonstrate convincingly each of these three points, given the results of Chapter 4, then I will have shown that the algorithm described by the Lewontin-Campbell computational theory applies to language evolution. The argument I am offering here is an “existence proof”; in other words, I am going to demonstrate that there exists linguistic phenomena for which the Lewontin-Campbell algorithmic theory obtains. I only wish to tackle an existence proof because, as you will see below, a general argument would require speculation into the role and activity of the text authors, and I wish to avoid such speculation.

5.3.1 Point #1 - individuals, traits, and heritability

1. A population of individuals over time with a correlation in heritable traits

Let's consider point #1. I stipulate that each chronicon describes a *population* and each text acts as the *individual*. The status of the individual in evolution has been the source of some attention and controversy (e.g., Hull, 1988; Ghiselin, 1997). And I hope it is not with too much circularity that David Hull argues that a “natural individual” is an individual on which laws of nature act (Hull, 1992). I am setting out to demonstrate that laws of nature, namely the Lewontin-Campbell computational theory, act on these texts as individuals. So this assignment at least stands ready to be refuted. But let me emphasize one thing: within the context of evolutionary models of language and culture, considering the text as the individual is really quite a radical twist. Without any exceptions (that I am aware of) within evo-

lutionary culture theories, the individual has always been the human actor (the author, potter, fisher, speaker, etc.) (e.g., Lumsden & Wilson, 1981; Cavalli-Sforza & Feldman, 1981; Boyd & Richerson, 1985; Barkow, 1989; Durham, 1991; Blackmore, 1999).

If the text acts as an individual then, over time, individuals are entering the population as they are published. The requirement described in point #1 is that there are groups within the population, lineages if you like, for which a correlation in traits exists across time. Note that there is no mention of any particular mechanism of heritability (Lewontin, 1970). It is sufficient that a group of individuals share traits relative to the population as a whole due to some history-minded mechanism (Dawkins, 1982).

To determine whether traits are shared across individuals we must decide what exactly is a trait. A "trait" is simply some observable feature of the individual. Colless (1985) identified three different semantics for "trait" (or character) in common usage within the biological community; Fristrup explains these nicely by analogy to eye color as:

- character-part - *Joe's blue eyes.*
 - character-variable - *Joe's eye color is blue.*
 - character-attribute - *Joe is blue-eyed.*
- (Fristrup, 1992, p. 46)

In other words, the trait can be thought of as the thing itself (the blue eyes), the category of thing (the color), or the attribute or value within the category (blue). Similarly, within the sentence "I see a bird," a trait could be:

- character-part - *The sentence's word "see."*
- character-variable - *A word in the sentence is "see."*
- character-attribute - *The sentence has the word-"see."*

While all three of these definitions for "trait" have found use within evolutionary theory, from the vantage of the Lewontin-Campbell computational theory they do not all make equal sense. Clearly, within this algorithm, a trait (should it exist) is that thing which is passed on (again, by some mechanism not stipulated) within the lineage. In organic evolution someone's blue eyes (character-part), the *Ding an sich*, are not passed within populations unless the individuals are undergoing optical transplants of some sort. Similarly, a category, such as eye color (character-variable), cannot be heritable in the sense meant by Lewontin-Campbell (sure, we all express, or don't, categories of traits, but the category as a conceptual variable is

not what is heritable). However, the values (character-attribute) are indeed passed on organically, e.g., my eye color is brown and so is my dad's.

Thus, for the Lewontin-Campbell computational theory a trait is an attribute of some relevant category. I believe all of the linguistic levels I studied in the previous chapter, lexical, lexico-syntactic, etc., are putative units of selection and thus relevant categories. (In Section 5.8 below, I will take up at some length this question of the size of a replicator.) However, for the current argument I'll consider only words as a relevant category. A trait of an individual within this population, then, is simply the presence or absence (more precisely the relative presence or absence) of the set of possible words — in other words a *lexical category-attribute*.

A side note: While eye color appears to us as a discrete trait (e.g., it is blue or brown or hazel), many traits are metric (e.g., your height). In my model the lexical category-attribute is metric insofar as I score individuals on the relative presence of words (or noun phrases, etc.) rather than a binary scoring of presence or absence.

Given this sense of "trait," in Chapter 4, I described sets of texts that shared many traits with each other relative to the population as a whole. These are the clusters, such as Clinton/Lewinsky and Guns. I know these texts share traits, as this is precisely the outcome of the clustering step described in Section 3.9. And, therefore, we have a population of individuals for which there is a correlation in traits across lineage (Dawkins, 1982; Lass, 1997).

We are left with one outstanding issue — Are these correlated traits inherited? In other words, are they the outcome of some historical mechanism of transmission? To answer this, it is important to understand what the alternative is, namely, convergent similarity. These correlations are either due to some mechanism of inheritance, or to chance events, or to some systematic process that produces the convergence.

The likelihood that these correlations are chance events is profoundly small: a few of them may be due to chance but the hundreds of correlations found in the clustering phases surely are not. The likelihood of specious correlations was analyzed in Section 3.11.

However, some of these correlated traits may be due to systematic convergence. In organic evolution, systematic convergence occurs as a result of environmental regularities that are similarly accommodated across non-hybridizing reticulate lineages. For instance, it is noted that many chordates respond to cooler climates with similar strategies, such as increased body size, decreased epidermal melanins, and so forth (e.g., Schmidt-Nielsen, 1983). These similarities across diverse species of chor-

dates are not always due to commonly inherited traits (e.g., birds and mammals). The distinction, then, is between homologies resulting from common descent and analogies due to other causes. A similar issue challenges historical linguistics' ability to resolve true cognates.

Since I have only taken on the task here of an existence proof, I simply need to offer examples of linguistic replicators, described in Chapter 4, that are due to inheritance instead of some systematic parallel convergence. At the lexical level these arguments seem the hardest to make. That notwithstanding, consider pejoratives such as "cunt" and "bastard" which occur approximately 100 times each in the Broaderick cluster of the Clinton chronicon (see Table 15). A substantial number of authors repeatedly made use of these same words. These words otherwise occur with relative infrequency, and are clearly not related to some common environmental regularity that might have drawn non-interacting authors to them, for instance, some aspect of the story seeding these exact word choices. For systematic convergence, one would have to posit some environmental regularity that would cause these authors, without reading or being influenced by prior posts to the thread, to make use of these identical pejoratives. And, further, these unrelated uses would have to be systematically correlated to a rise in the volume of published texts on a similar topic. This strikes me as fabulously unlikely. It is worth repeating: no particular mechanism of transmission is required or implied. This transmission could be mediated through a rich and varied cultural nexus.

I believe similar arguments can be made for many of the words found in other clusters (e.g., "moral," "public," and "denial" in the Clinton/Lewinsky cluster). And I believe this is even more obviously true when larger replicator types are considered, such as phrases, where the chance of convergence not due to any heritable transmission is greatly reduced (e.g., "his personal life," "immediate threat to your life," or, "Clintonphobe grind tooth").

I claim that for point #1, the existence proof is supported. However, exactly how to determine, in all cases, when we have convergent analogies versus homologies due to common descent (versus the occasional chance event) is an important open question.

5.3.2 Point #2 - variation

2. Expressed variation

Point #2 requires that individuals in the population be different. In particular, these differences should be expressed by the individual in such a way that the evolutionary algorithm has access to them. Clearly, lexical character-attributes, the relative presence of words in texts, are highly variable. Moreover, these variations are expressed in the “morphology” of the published text. If the word-”see” trait is high, then we will be able to observe a relatively large number of occurrences of “see” in the text. The evolutionary algorithm operates *on* the entire text, where these variations are visible, but *for* the particular trait. This presence of variation in word usage is enough to confirm an expressed variation of traits.

5.3.3 Point #3 - trait/fitness covariance

3. *A covariance between variants and the success in time of the population*

The final step to the Lewontin-Campbell computational theory of evolution states that there must be a covariance between the variant forms of heritable traits within a lineage and the relative abundance, or fecundity, of that lineage. In other words, the traits must covary with the replicative success of the individuals that possess them. I will refer to this as a trait/fitness covariance.

Fitness (at least in this context) is a statistical abstraction (Williams, 1966). It is a summarization for a population rather than a property of some individual. This is different from an individuals’ Darwinian (or inclusive) fitness, which is the individuals’ expected contribution of offspring to the next generation (as well as those of its kin) (Hamilton, 1964; Williams, 1966; Durham, 1991).

For our purposes, an evaluation of the summarizing *average population fitness* is sufficient, and this is given by the lineage population size over time (Crow & Kimura, 1970). This measure is suitable since, to establish a trait/fitness covariance, we are not interested in whether some individual is fit, but whether certain traits contribute to the rise or fall of fitness, on average, across the lineage.

This correlation, between the variance of a trait and the average population fitness, is precisely the measurement I’ve been making in order to establish replicators as active. “Active replicator” is just a different way to name the trait/fitness covariance of the Campbell-Lewontin algorithm, a name that is designed to focus our attention more squarely on the unit of replication (Dawkins, 1982).

Importantly, this formulation saves the Microevolutionary Language Theory from being a tautology (as is occasionally claimed of socio-cultural models of evolution) as it defines fitness as separate from that which evolves. It is at the level of replicating traits that we expect accumulating design due to evolution. Yet fitness is defined in terms of survival, and replication of texts, as individuals in the socio-cultural milieu. If the Microevolutionary Language Theory merely claimed that “To be or not to be” is a powerful language replicator because it is common, then it truly would be a tautology.

5.3.4 *Quod Erat Demonstrandum*

This reading of the results of Chapter 4 argues that the Lewontin-Campbell computational theory of general evolution holds for the evolutionary language model I’ve developed. To demonstrate this, I posit that each of the requirements of the Lewontin-Campbell algorithm obtains — replication and heritability of traits, variance of those same traits, and covariance of the traits with population fitness. Let me ground this in an example, the word “Clintonphobe” appears with variation over time in a collection of NetNews posts dealing with the Clinton/Lewinsky scandal. This trait does not reappear over time due to chance convergence, but instead it reappears due to some history-minded mechanisms of heritable transmission. The presence of this trait (in an SVO lexico-syntactic replicator) covaries with the average population fitness for this population of texts. The more this trait is expressed, the more texts appear within this population.

The hypothesis that language (at least at this micro-level) executes the Lewontin-Campbell algorithm, I am calling the Microevolutionary Language Corollary. This corollary (see Figure 51) can be simply stated as:

Microevolutionary Language Corollary: Evolution of language and organic evolution are isonyms of each other (and hyponyms to a general process of evolution).

5.4 Hull-Dawkins Typological Theory

I believe the Lewontin-Campbell computational theory is the most powerful model of general evolution we have. But there exists one other model in strong competition which I am calling the *Hull-Dawkins typological theory*. This name, again, borrows somewhat from Plotkin (1994), as well as from those who have obtusely referred to this model as the Hull-Dawkins distinction (e.g., Eldredge, 1989; J.S.

Wilkins, 1998). I call this a *typological theory* because it proposes a set of constant diagnostic characteristics that accompany any evolving system.

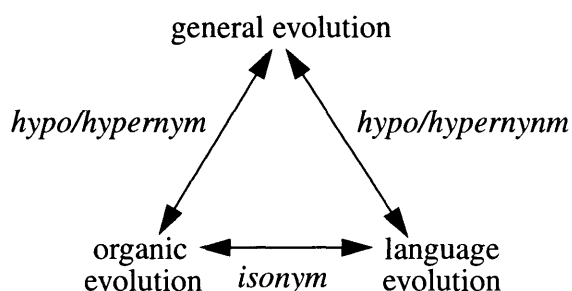


FIGURE 51. Microevolutionary Language Corollary: organic evolution and language evolution are isonyms of one another.

I will here discuss the active language replicator model in the context of the Hull-Dawkins typological theory for two reasons: First, the theory is certainly important, and its explanatory powers are worth testing against my model (and vice versa). Second, the theory is particularly useful in establishing a replicator-eyed viewpoint of evolution; that is, looking at evolution from the vantage of the replicator, which will help to illuminate some of the novel explanatory avenues revealed by an active language replicator model.

The first and principal component to the Hull-Dawkins ontology we already know, namely, the *active replicator* (Dawkins, 1976, 1978, 1982). For Dawkins, a replicator is “anything in the universe of which copies are made.... [and an] *active replicator* is any replicator whose nature has some influence over its probability of being copied,” (Dawkins, 1982, p. 83, emphasis in original). What Dawkins is most keenly interested in are *active germ-line replicators*; these are active replicators that are part of a line of descendant. These concepts are, of course, central to the activities described in Chapter 4; I set out to discover through computational and statistical means active germ-line replicators within natural language.

Dawkins realized that replicators, by and large, are not directly selected by the evolutionary process. In fact, replicators mostly ride around in other entities (for instance organisms) and selection acts on these agents. Dawkins referred to the

entities that conglomerate replicators as *vehicles*, noting that a “vehicle is any unit discrete enough to seem worth naming which houses a collection of replicators and which works as a unit for the preservation and propagation of those replicators” (Dawkins, 1982, p. 114).

David Hull (1980), in an influential paper, noted that Dawkins’ concept of replicator and vehicle confused two separate qualities of replication. First, replicators influence their environment by being party to the production of their copies. But, further, they influence the environment by impacting the expression of features through a phenotypic effect. It is the much more active sense of replicators with phenotypic influence that Hull calls an *interactor*. Hull argues that an interactor is “an entity that directly interacts as a cohesive whole with its environment in such a way that replication is differential,” (Hull, 1980, p. 318).

In addition to interactor, Hull added one other concept to the ontological mix: the *lineage*. Hull claims that “[replicators and interactors] are the entities that function in the evolutionary process. Other entities evolve *as a result* of this process” (Hull, 1980, p. 327, emphasis in original); it is the lineage, he then argues, that evolves as a result of the evolutionary process.

It is helpful to ground this in an organic example: A replicator is a gene. An interactor is the organism. And a lineage, the species. Williams argues persuasively that even though this example is common, we need to remain mindful that the Hull-Dawkins typological theory is general across substrate materials and indeed true for pure information (the codical domain) (Williams, 1992, p. 10).

Thus, the Hull-Dawkins typological theory of general evolution posits three main entities — the replicator, the interactor, the lineage — as the necessary and sufficient components to an evolutionary process. Identifying these components is enough to reveal an evolutionary process. Plotkin summarizes this by stating, “if entities that can make copies of themselves (*replicators*) are propagated in space and conserved in time because of the differential extinction and proliferation of *interactors*, these will in turn lead to historical changes in *lineages* and evolution will have occurred” (1994, p. 97, emphasis added).

5.4.1 Hull-Dawkins meets language replicators

Given this typological theory, how might it apply to my model of text evolution? The concepts of replicator and active replicator have been treated *ad nauseam* in the previous chapters, and I claim to have identified active language replicators at mul-

tiple levels. In Section 5.3.1, I considered lineages within the chronica and argued that clusters make up a lineage because they describe a plexus of individuals that share history-minded traits relative to the population at whole. So, for example, I showed that “his personal life” is an active lexico-syntactic replicator within the Clinton/Lewinsky lineage.

Identifying vehicles within my model, should we wish to, is a pretty straight-forward business. Recall that vehicles are objects, potentially quite static, that house collections of replicators. The individual texts making up each chronicon fit the definition of vehicle quite nicely. They express collections of replicators, or at least the phenotypic effect of replicators, and are involved in the differential transmission of traits and the appearance of subsequent texts.

David Hull criticized the definition of vehicle because it was too passive and static. His notion of interactor suggests a much more dynamic object, an entity that sits quite directly in the causal pathways of differential selection and replication. Could the texts themselves serve the function of a dynamic interactor? Intuition might suggest that texts are the acme of passive, static objects.

In considering this it may be helpful to ponder an organic example of a fairly static interactor that is able to manipulate its host’s behavior. Dawkins (1982) has discussed the interesting case of the fluke or “brainworm” (*Dicrocoelium dendriticum*) (Wickler, 1976; Love, 1980). This nasty critter burrows itself into the brain of an ant. By entering its suboesophagela ganglion, the ant is compelled to climb to the top of blades of grass, clamp on, and remain stationary. This behavior puts the ant at great risk of being eaten by grazing sheep — which is, in fact, the exact outcome the fluke is hoping for. The ultimate target host of the fluke are ungulates, and by modifying the ant’s behavior it seeks to accomplish its goal of taking up cozy residence in a sheep.

I mention this story as analogy to one way in which texts might be considered active interactors. If a human reads a particular text, then perhaps the text enters the mind of the reader in a way analogous to the worm entering the brain of the ant. And this text can actively alter the human’s behavior, “causing” it to author (or not) texts that might share certain traits with the original text.

This model of texts as parasitic interactors is similar to the *virus of the mind* (Dawkins, 1993) framework of cultural evolution. It posits that ideas parasitize human brains and potentially modify their host’s behavior. There has been some successful work in “idea virology” (e.g., Marsden, 1998). The general concept, that some ideas may spread epidemiologically, seems a fruitful avenue for exploration. How-

ever, this concept, and in particular the analogies to parasites, has mostly been usurped by a collection of pop-scientists (Brodie, 1996; Lynch, 1996) whose arguments run to the perverse extreme (see Marsden, 1999 for a critique). However, independent of the quality of this work, I find myself genuinely uncomfortable with the overall idea — that human’s are, somehow, unwitting hosts to ideas which parasitize them, driving their behavior.

I believe it is possible to save the active language replicator model from the ranks of the mind virus, and yet still make use of Hull’s interactor concept, by expanding what I consider the interactor beyond simply the text. Why not consider the interactor as both the text and the human who reads it (or some aspect of the human’s cognitive psychology)? In other words, the interactor is *text + human psychology* in rich interaction. This entity is surely an active one that is able to express traits and to directly participate in processes of differential selection. Happily, it places the human back into the drivers’ seat; it does not characterize all ideas as viruses that parasitize human minds. In Table 31 I map my model to the Hull-Dawkins typological theory (cf. Sereno (1991) for a different set of correspondences):

Hull-Dawkins theory	Microevolutionary Language Theory
individual	published text
population	chronicon
replicator	lexical, lexical co-occurrence, lexico-syntactic, or syntactic replicators
active replicators	replicators with trait/fitness covariance
vehicle	published text
interactor	published text + human psychology
lineage	cluster of published texts

TABLE 31. Mapping from the Hull-Dawkins typological theory to active language replicator model of Microevolutionary Language Theory.

These distinctions, between vehicle and interactor, are subtle and difficult issues in the Hull-Dawkins theory. This is a primary reason why I favor the Lewontin-Campbell computational theory to this typological description of general evolution, especially since in all likelihood, they both describe an identical process (Plotkin, 1994). But what the Hull-Dawkins theory has done, to its supreme credit, is to convincingly resolve a fundamental question of the evolutionary process: what is the

central beneficiary of evolution, in other words, who does evolution act *for*? An attention to this question (and answer), within the context of my model of language evolution, provides some new insights. It is these issues I will take up in the next section.

5.5 Units of Selection

The Hull-Dawkins typological theory sprung from a discussion on the *units of selection*. This question has had more than a few formulations (Lloyd, 1992). A particularly useful one is this: For whose benefit is evolution? This is a question that is so important, it is worth saying in Latin as lawyers do: *Cui bono* (Dennett, 1995)? Asking this question is the same as asking at what level *adaptations*, or special and complex functional design, will accrue (Williams, 1966; Maynard Smith, 1972; Dawkins, 1983; Eldredge, 1989; Pinker, 1997). (In Section 5.6, I will discuss exactly what an adaptation is.) I believe this enterprise, determining the central beneficiary of the microevolutionary process in language, opens up new explanatory pathways for historical linguistics, language change theories, and by extension, social anthropology. To understand the answer (and, indeed, the question) it is useful first to gloss the history of this debate on units of selection.

5.5.1 Historical review

A glance across the last 200 years of biological thought shows that a number of entities have been crowned the central beneficiaries of the evolutionary process. Early evolutionist (e.g., Lyell, 1863) centered the process on the species. Enter Darwin, with one of his most radical and important contributions; he asked not what is good for the *species*, but what is good for the *individual* (Mayr, 1991). To recast this into the Hull-Dawkins universal ontology — the question for Darwin was not what is good for the lineage, but what is good for the interactor. This reorientation is part of Darwin's greatest theoretical achievement, namely, *population thinking* (Mayr, 1991; Dennett, 1995). Population thinking recognizes the singular importance to the evolutionary process of variation between individuals within a population. Prior to this reformulation, species were seen as immutable, Platonic, "natural kinds." Darwin recognized the uniqueness of individuals. He reasoned that if all individuals within a species are not identical, then a benefit to one need not accrue to all.

This focus on the individual lasted for quite awhile. But an influential book, written by V.C. Wynne-Edwards (1962), shifted focus away from the good of the individual to the good of the group. Wynne-Edwards set out to explain altruistic behaviors that

were of benefit to a group, seemingly at the expense of the individual. However, this approach was short-lived and fell at the hands of a number of powerful arguments, which noted in particular that group selection could not be stable in the presence of individual defectors. In other words, group selection did not offer an evolutionarily plausible explanation.

Principal amongst these arguments against group selection was G.C. Williams' (1966) *Adaptation and Natural Selection* — the book that, to this day, remains the most important piece of writing on evolution since Darwin. Williams shifted focus away from the group, but he did so without setting it back only to the individual. He argued that whatever the answer to the *Cui bono?* question, it must certainly be something that is long-lived. The essence of evolution is the production of variation and the differential selection based on these variants, resulting in the accumulation of adaptation. This fundamentally is a process of many generations, and *Cui bono?* is, in the end, an actuarial question.

Williams develops this line of thought by evoking the ghost of Socrates. Socrates' body is long gone — the sorry outcome of dining on hemlock. Williams notes that while “natural selection may have been acting on Greek phenotypes in the fourth century B.C., it did not of itself produce any cumulative effect” (1966, p. 23). For all we know, Socrates' hereditary line may well have ended. But whether it has or not, certainly his genotype is no longer with us. Williams concludes that the “loss of Socrates' genotype is not assuaged by any consideration of how prolifically he may have reproduced. Socrates' *genes* may be with us yet, but not his genotype, because meiosis and recombination destroy genotypes as surely as death” (1966, p. 24, emphasis added).

With this, Williams is arguing that neither the individual nor the lineage nor the genotype have enough permanence to stand as the beneficiary of the evolutionary process. In contrast, the *replicator* (in this case the gene) is highly conserved in time, and therefore, does indeed stand to benefit. This replicator orientation was further elucidated by, in particular, Dawkins (1976). For him, the replicator as central beneficiary to the evolutionary process remains the “primary fact about evolution” (Lloyd, 1992, p. 337). Natural selection acts *on* the interactor but selects *for* the replicator. This viewpoint now stands as a conventional and orthodox cornerstone to contemporary evolutionary theory (Plotkin, 1994; Pinker, 1997).

In summary, a principal outcome of Williams' arguments and the formulations of the Hull-Dawkins typological theory of general evolution is the interrogative pairing:

Question: *Cui bono?*

Answer: The active replicator.

5.5.2 *Cui bono?* in text

Williams, Hull, Dawkins, and others established the answer to the *Cui bono?* question in general. I believe that asking this question of the active language replicator model is of real value. As we shall see, this orientation represents a novel and potentially useful alternative to a range of current structuralist thought in the social sciences.

Williams has noted that Socrates' genes remain largely with us today and highly conserved in time. It is unfortunate for me that Williams chose Socrates to make his point; purportedly, Socrates did not write anything (*Encyclopædia Britannica Online*, 1999)! But that notwithstanding, I claim that Socrates' words are with us to this day as well, they, too, being highly conserved in time.

I established in the previous chapters that there are certain active language replicators which I was able to distill from a variety of text collections. The Hull-Dawkins typological theory claims that these active replicators are units of selection within an evolutionary process; thus they sit as the central beneficiary to that process, and it is at this level that adaptations should accrue (Williams, 1966; Maynard Smith, 1976; Eldredge, 1978). We expect functional complex design at the level of the replicator. Thus, the interrogative pair for this model is:

Question: *Cui bono* in natural language?

Answer: The active language replicator.

This is a fundamental outcome of the replicator-eyed orientation and is the core "truth" to a Microevolutionary Language Theory.

5.5.3 *Cui bono?* as explanation

That simple language replicators, such as words or phrases, might be the central beneficiaries of the evolutionary process within natural language is a startling observation that arises out of the conceptual integration of contemporary evolutionary theory with corpuslinguistic models of language use. The possibility that *adaptations* accrue at this simple level is particularly striking. This expectation is in contrast with special design at the level of the text (e.g., *Moby Dick*), the language (e.g., English), the author (e.g., Joseph Conrad), and so forth. Intuition or, indeed,

experience suggests placing design and benefit at the feet of these other units and not with things such as words and phrases.

This unexpected outcome offers new explanatory avenues that I believe have, by in large, not been previously available. It allows us to explain a socio-cultural phenomenon, in this case a linguistic one, by considering the good of these simple, replicating entities.

In Section 5.6, I will employ this finding while exploring special design in lexical semantic pejoration. However, first let me illustrate the potential of this replicator-eyed orientation with a “my summer vacation” story that will no doubt have a familiar feel.

A few years ago I was traveling within the Dogon region of Mali in the Sahel of West Africa. I was two days by foot away from electrification, plumbing, and car-bearing roads. In fact I was quite literally half-way to Tombouctou! One night, as I sat with friends and the chief of a local village, the quiet dusk was broken by the raising of a few young voices:

Un, Dos, Tres! Ole, Ole, Ole!
Un, Deux, Trois! Ale, Ale, Ale!
Here we go! Ale, Ale, Ale!
Go, go, go! Ale, Ale, Ale!

It was a group of Dogon kids singing Ricky Martin’s “The Cup of Life.”

I said this story would have a familiar air, and it does. We all know, and many of us have seen first hand, that no matter how remotely you travel many culture elements have beat you there. But let’s make no mistake — something here demands an explanation. Why did this snippet of language travel so far with such fidelity and fecundity? What advantage does the singing of this song deliver to those Dogon kids? their community? Ricky Martin?

These are not novel questions; this is the bread-and-butter for many programs of social and cultural anthropology. And there is nothing new to applying biological or evolutionary theories to the study of such cultural and linguistic processes (Kuper, 1999). But there is something novel in applying the Hull-Dawkins typological theory and a replicator-eyed viewpoint to these questions. In other words, the Micro-evolutionary Language Theory suggests that the answer to *Cui bono?* could be at the simplest level of language replicator. Perhaps “Ole, Ole, Ole!” sits as the central beneficiary of this cultural evolutionary process and from the vantage of the evolu-

tionary process nothing else need benefit (nor be harmed, for that matter). Not the kids, not the community, not Ricky Martin.

This explanatory avenue is novel and has not been available to the current range of functional and structuralist thinking within social and cultural anthropology (Aunger, 1999).

5.6 Microevolution and Complex Design

I will now take this notion of a replicator-eyed view a step further and argue that a particular lexical replicator is an adaptation and has accrued special and complex design. This is the heart of the Microevolutionary Language Theory. But, first, let's look more generally at this notion of "microevolution." I've referred to this work as a microevolutionary theory of text. However, so far I have not said much about what that means, leaving it to intuition or prior knowledge, that this refers, in particular, to small scale changes. Indeed, microevolution does refer to small, gradual, and vertical (temporal) change (Durham, 1991; Plotkin, 1994), in other words, *natura non facit saltum*. It is now reasonable to consider in more detail what that means and, at least, attempt to verify that it holds in our linguistic analysis.

The graph in Figure 52 shows the relative presence, normalized for volume of texts and with the misleading zero-points removed, of the "Monica S. Lewinsky" trait

within the Clinton/Lewinsky cluster of the Globe chronicon. This is a portion of the trait as graphed in Figure 8.

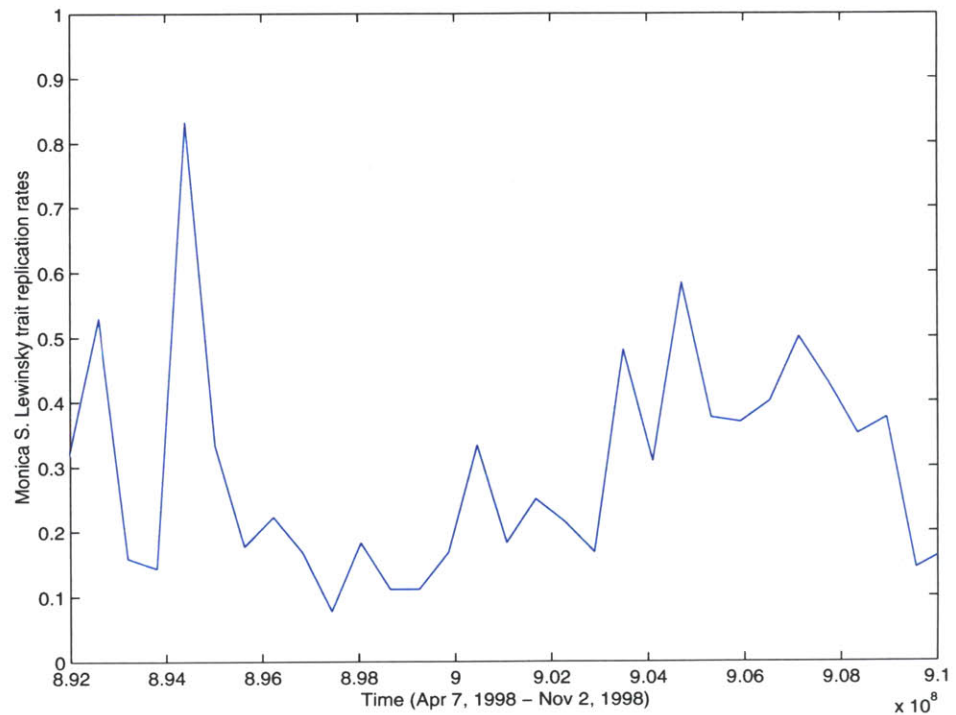


FIGURE 52. The ‘Monica S. Lewinsky’ trait - vertical, small, and gradual.

First, we know the variation is vertical insofar as it is across time; this is analytically true. But what does it mean for the variation to be “small?” Small relative to what? The largest change is from approximately 0.2 to 0.8 in the course of one week (seven generations for the news). What is certain is that there is nothing so absurd that it can not be found somewhere in nature (Hull, 1982); a times-four variation in some trait across a collection of generations hardly seems unreasonably “large.”

But the question of gradualism is more important. And I believe that this graph is not enough to determine if the variation is gradual. Gradualism is a property of the population, and not any individual within the population (Mayr, 1991; Plotkin, 1994; Dennett, 1995). For the appearance of this trait to be “intergradational” (Sim-

pson, 1970) it is necessary to establish that the constituents of the trait (e.g., the name “Monica”) were present as gradual steps, without jumps, within the population at large and available as an innovation. This can be true, even though Monica S. Lewinsky veritably jumped onto the stage of public discourse when her story was broken by Matt Drudge. Thus, even a sudden rise in the trait is still gradual innovation when considering the population as a whole.

5.6.1 Adaptive value and adaptation

Microevolution does mean small, gradual, and vertical change. But, most importantly, it refers to change that is generally adaptive or, indeed, an adaptation (Simpson, 1944; Mayr, 1991). In Section 5.5, I argued that because we have active replication the Hull-Dawkins typological theory tells us to expect adaptations to accrue at the level of the replicator. I will now explore this notion of adaptation, and see if it opens up any new explanatory pathways.

The idea of *adaptation* is the most enduring and powerful concept within evolutionary biology (Plotkin, 1994; Dennett, 1995) and it is fraught with controversy and confusion. As Williams famously put it, “adaptation is a special and onerous concept that should be used only where it is really necessary” (1966, p. 4; see also Gould & Lewontin, 1979). And Greenberg (1992) has noted that if the problem of adaptation is difficult in organic biology, it is even more difficult in linguistics.

An adaptation is some feature of an individual that helps it survive or reproduce. Identifying some trait as an adaptation begins with a demonstration of its adaptive value *vis* its trait/fitness covariance; the possession of the trait must correlate with the replication success of the trait’s possessor (Reeve & Sherman, 1993). If such a covariance is shared by unrelated groups who are responding to similar selective pressures, then evidence of an adaptation mounts (Lewontin, 1978). To some, this ahistorical demonstration of fitness enhancement is enough to label the trait an adaptation (Clutton-Brock & Harvey, 1979; Reeve & Sherman, 1993; but compare Gould, 1984). For others (Williams, 1966; Lewontin, 1978; Sober, 1984; Waddington, 1957; but compare Bock, 1980), a history-laden investigation of the trait is critical. In particular, the trait must have established itself in trans-generational time due to some design quality it possess relative to variant forms. Understanding this historical form of adaptation is primarily a project in reverse-engineering the complex functional design elements of a trait (Dennett, 1995; Pinker, 1997).

The particular trait I have most closely studied as an adaptation is the “Nazi” replicator, which I described at length in Section 4.3.3. Recall that I discovered lexical

co-occurrences of the pejorative use of “Nazi” within three diverse online settings — skeptical discussions of science, debates on the U.S. Constitution, and sexual fetishism. I discovered in all three of these environments that the possession of this trait covaried with an increase in average population fitness of the lineage — a trait/fitness covariance. That each of these occurrences of the trait admitted to a trait/fitness covariance convincingly establishes that the trait is of adaptive value (Reeve & Sherman, 1993). And that the same trait is of adaptive value across multiple environmental settings adds strength to the claim that it is an adaptation (Lewontin, 1978).

5.6.2 An historical demonstration of adaptation

I’m claiming that the use of “Nazi” as a name-calling device is strongly adapted, since I found it to be of high adaptive value within multiple groups of texts. But to claim that this use of “Nazi” is an adaptation requires linking the trait to its history (Waddington, 1957; Williams, 1966; Lewontin, 1978; Sober, 1984). Have there been variants in usage of “Nazi,” from which this particular pejorative name-calling usage has been selected and developed over generations, due to some design quality? For Williams, this question is the *raison d’être* for his research programme. In his words, “the central biological problem is not survival as such, but design for survival” (Williams, 1966, p. 159).

To establish the trans-generational selection for this word-usage, let’s consider the semantic variants of “Nazi” over time, and ask how socio-cultural forces have described a selective environment in which the pejorative usage has thrived. This requires a bit of just-so story telling: that is, it amounts to a plausible bit of reverse-engineering that stands ready for further testing.

To understand this long-term usage history, it is not enough to look at recent usage within the NetNews corpora; instead, it requires study of the word’s semantic change over time. A standard jumping-off point for such a study begins with the *Oxford English Dictionary*, which gives us a collection of word-usage variants. “Nazi” originally appeared in 1930, as a short-hand reference to the National-Socialist Party in Germany (Simpson & Weiner, 1989). By 1949, there were word-meaning variants in which “Nazi” is used to describe any “political organization with similar aims, beliefs, or methods” of the German National-Socialists. And by 1973, “‘Nazi’ has become an indiscriminate political cliché.” More recently, the set of word-meaning variants has included “Nazi” as a fairly generic pejorative name-calling device with no particular reference to the German National-Socialists nor,

for that matter, to politics at all (Maas, 1994). Thus, there have been and continue to be live semantic variants of “Nazi.”

What does this list of variant word-usage tell us about the semantic history for “Nazi?” Semantic change is defined, generally, as a change in the context in which a word might appear (Jeffers & Lehiste, 1979). Semantic innovation does not produce a total change in meaning, as much as an additional meaning variant. It is in a subsequent semantic change, where some word meaning is removed from use, that the possibility for an absolute semantic change exists (Brown, 1979; Wilkins, 1996; Lass, 1997). While general laws of semantic change have been elusive (Anttila, 1989), a number of regular patterns have been described (Brown, 1979; Jeffers & Lehiste, 1979; Traugott, 1985; Wilkins, 1996) which include generalization (a word meaning is extended to cover more cases), specialization (a word meaning is narrowed), metaphor (a word meaning is transferred to a new referent, suggesting a similarity), and metonymy (identifying a whole by its part). Less frequent are pejoration (a word develops a more negative meaning) and amelioration (a word develops a more positive meaning) (Traugott, 1985).

In Figure 53, I gloss the major semantic changes which have occurred for “Nazi.” The first change is one of simple generalization. The example from 1943 shows how the term “Nazi” was expanded to include other political parties and groups which had similar aims or methods of the German National-Socialists. Indeed in the 1940’s these groups were often directly linked with their German counterpart. This process of generalization continued to the current common use of “neo-Nazi” as a referent to any group, perhaps only loosely organized and political, that shares the original aims or methods of the German National-Socialists. The second major

semantic change was the process of pejoration in which the word came to act as a general attack word. We will examine this variant at some length.

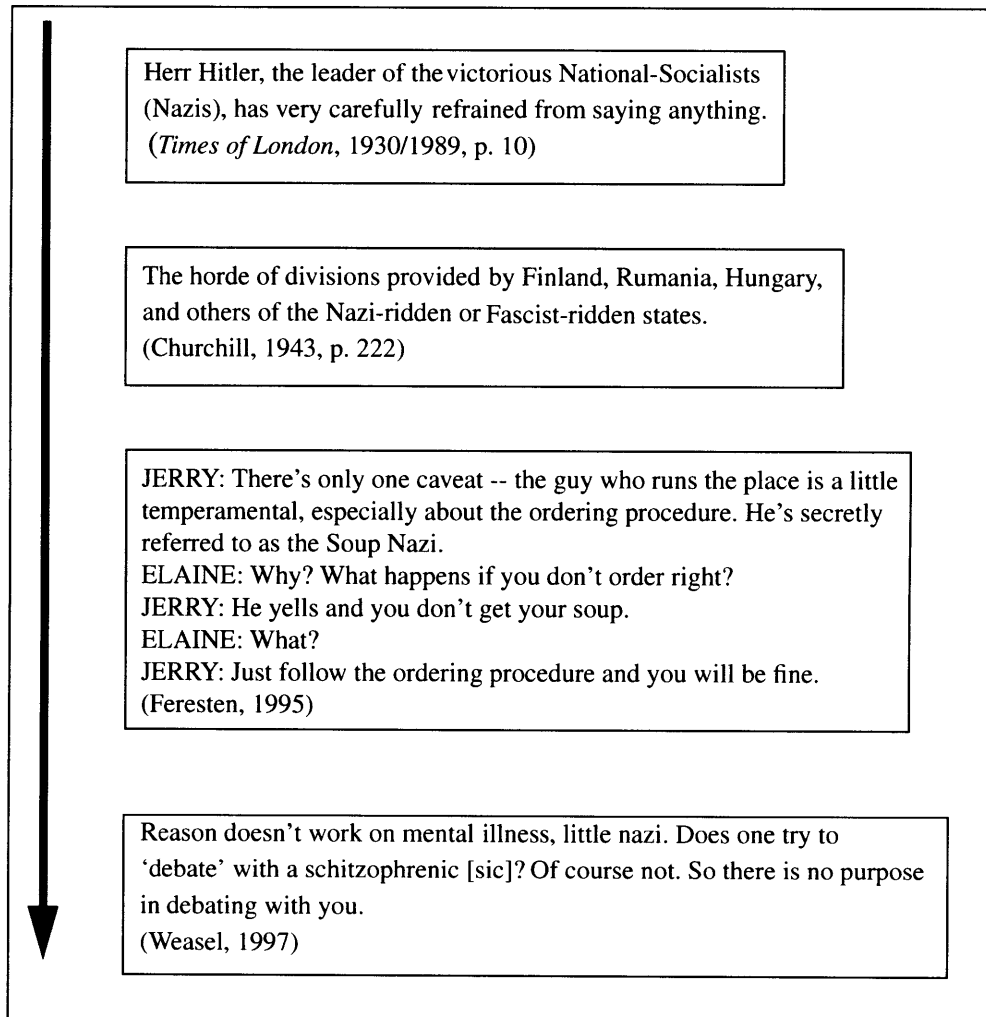


FIGURE 53. Major lexical semantic innovations for "Nazi."

5.6.3 Pejoration and selective forces

The second to bottom quote in Figure 53 is from a well known episode of *Seinfeld* in which a severe and strict purveyor of soups is referred to as “Soup Nazi.” The bottom quote is from a NetNews text in the *Skeptic* chronicon. It shows the use of the word “nazi,” now decapitalized, as a general pejorative attack name. In both these examples, no reference is made to German National-Socialists nor to politics. The general sense of this word-usage is that the person described is stubborn, strict, and unmoving. This process of pejoration into a clichéd hypocoristic attack name is an example of an often successful word-meaning substitution (Kleparski, 1986).

My goal is to reverse-engineer this pattern of semantic innovations in order to establish that this particular word-usage variant was selected due to some design feature it possesses. We know that the selective forces operating on innovations within language are a function, primarily, of the socio-cultural environment; all of language change has at least some socio-cultural component (Salmons, 1990). But semantic change, more than any other language change (e.g., morphosyntactic change), is linked with socio-cultural causes (Arlotto, 1972). I wish to establish that this process of pejoration came about due to *selection*, as in the Lewontin-Campbell evolutionary algorithm, on these variant forms. To establish selection as the prevailing force, I need to rule out other alternative explanations. The other potential forces for semantic change are *chance* accumulation of innovation and strictly *cognitive* (or psychological) forces, such as ease of pronunciation or recall. These mirror similar potential causes of genetic change in organic evolution — drift and environmental forces (Pinker & Bloom, 1990). These other forces will be taken up again in Section 5.7.

The simplest way to rule out random chance as the source of these semantic variants is to appeal to their complex nature. One might find cases of simple linguistic change, for instance spelling variation, that are strictly due to chance events, such as copy errors. But if random drift is to account for this pejoration, the word would have to be stable enough in these new contexts so as to accumulate usage. However, a complex semantic shift of this sort is not expected to be interpretable by language users in the absence of either an appropriate socio-cultural or cognitive environment (or both) (Brown, 1979). Thus, chance accumulation of such usage in the absence of socio-cultural or cognitive forces is most unlikely, since it would amount to usage without understanding.

The reason that the pejorative innovation is interpretable by English language users is due to a regularity in the information-processing environment, to wit, the strong dislike of German National-Socialists. This regularity could not be due to a strictly

cognitive mechanism. While humans may have evolved mechanisms from Pleistocene conditions (in fact probably did) which contribute towards hatred of other groups, it is impossible to conceive of how humans might have evolved cognitive mechanisms to hate the *particular* group of German National-Socialists. Thus, it is the socio-cultural transmission of this dislike which creates a selective environment sympathetic to the pejorative process. While cognitive forces may well have contributed to the process of pejoration, socio-cultural forces seem to have played the major role. In either case, however, it becomes clear that the pejorative variant was established via evolutionarily significant selection forces, and how, exactly, it divides between socio-cultural and cognitive mechanisms is not what is at issue.

To say that the socio-cultural forces produced a selective environment favorable to the pejorative attack variant of “Nazi” is also to say that the variant is able to solve some specific, adaptive problem posed by the information processing environment (Tooby & Cosmides, 1992). That the word-usage solves some adaptive problem requires argument from a particular and peculiar vantage — identifying the functional qualities of the word-usage as somewhat separate from the word’s ideational or iconic meaning. That is, we must reverse-engineer its function. For many adaptations within the natural world, we are able to reverse-engineer their function by use of direct evidence or simple observation. Just so, we are able to observe the properties of the pejorative use of “Nazi.” What happens when people are called “Nazi” in a dialogue without reference to German National-Socialists? A simple observation of the word-usage indicates that it angers and excites its audience. We can *observe* the word-usage solving the problem of audience stimulation.

Consider a counter-example of a potential semantic change that has, so far, not occurred. The word “Scout”, as in a “Boy Scout,” may find itself, through the process of amelioration, coming to stand for individuals or groups who are well-prepared, gracious, help old ladies to cross the street, and so forth. Or the word “Scout” may come, through a process of pejoration, to stand for individuals who are anti-gay and anti-atheist, since the Scouting movement is currently embroiled in controversy around its treatment of gays and atheists. But if one attempted today, amongst English users in the U.S., to use the word “Scout” in the context of an anti-atheist group, this innovation would not be readily interpretable. The current culturally selective environment does not favor (apparently) this word-usage.

Let me now summarize: The occurrence of “Nazi” as a pejorative attack name has strong adaptive value, as demonstrated by its large trait/fitness covariance among multiple groups within NetNews. I demonstrated through computational microevolutionary analysis that as the particular word-usage becomes more salient the vol-

ume of texts within that cluster increases. This was shown to be true in three different collections of texts dealing with three fairly different sets of subject areas.

The particular design feature, “Nazi” as a general purpose attack word, has developed as part of a larger process of pejoration as demonstrated by its linguistic history. I have argued that the variant has installed itself across trans-generational time due to the selection on some complex functional design feature. The variant did not arrive because of drift, nor simply because of cognitive factors. The variant exploits language users strong negative reaction to German National-Socialists, a regularity in the socio-cultural information processing environment. Moreover, it solves a particular adaptive problem, it excites, and often angers, its audience.

By demonstrating this trait’s adaptive value in the present, and by linking the trait to selective forces in its past, I’ve given evidence that this trait is an adaptation and an example of complex functional design at the simplest level within English.

5.7 Selection as a Strong Force

The most important and novel contribution of Darwin was his theory of *natural selection* (Mayr, 1991). In fact, many people think of selection as the entirety of the Darwinian Theory. In the previous section I argued that certain forces select for the “Nazi” trait. However, I have yet to firmly establish that selection is, indeed, a strong force behind the dynamics of natural language.

Peter Grant argues that “the essence of selection is that certain individuals in a population do better than others in part because they possess traits, or expressions of traits, not possessed by other individuals” (Grant, 1986, p. 184). Thus, selection is the mechanism by which individuals differentially reproduce over time. In this section, first I will argue that selection, in contrast to direct engineering, is the primary explanation for complex design within language. Then I will review, and reject, nonselectionist forces that have been advanced within organic evolution as potential explanations for complex design.

I have observed empirically the differential survival of variant traits within the chronica, and reported extensively on those observations. And further, I have claimed to discover functionally adaptive and complex traits, such as the “Nazi” trait, above, and have linked them to selective forces operating over time.

When we observe functionally complex objects in the world, we assume a designer has been at work. This, of course, sits at the core of William Paley’s (1803) famous

argument: the existence of a watch presupposes a watchmaker. While we discount the existence of an invisible engineer within the natural world, at first blush, this may seem like an appealing and available solution to the language design problem. In other words, why can't the existence of complex language traits at the simplest levels be the work of a collection of *languagemakers* — all of us who use language? This might indeed provide a more parsimonious solution than relying on evolution for complex design.

All language users are “languagemakers” of some skill; for instance, each contributor to the *Globe* chronicon is an expert engineer of language complexity. And a description of the texts published across time clearly provides a diachronic and historical review of rich, language engineering activity. It does not, however, give a theory of the establishment of complexity *within* the language that the authors find themselves using. This is something altogether different. By analogy, you can compare the skills and fluency that *Boston Globe* reporters have with their computer publishing tools. However, this is *not* a theory of software design.

Why is an explanation of languagemaking, by the authors of articles published within *The Boston Globe*, not sufficient to explain functional complexity within the English language? Because it only gives half the picture. The authors contributing texts to the *Globe* chronicon are engines of *variation* within the population. But they do not, as collections of individual authors, act as *uber* engineers over the selective retention of variants (some of which they indeed may contribute). How could they? To assume otherwise would give them “invisible hand” powers over their socio-cultural milieu.

In other words, I am stressing the clear de-coupling between the individual engineering of variants within some population and the selective retention and reoccurrence of those variants, trans-generationally, within the same population.

It is no doubt true that these variants are often engineered with a clear eye toward success within the selective field they operate. For instance, authors to *The Boston Globe* operate intentionally, engineering texts they think have a high probability of success within the socio-cultural landscape. Variation is clearly not random and this might seem contrary to the evolutionary algorithm. But random variation is neither necessary nor generally observed within evolving systems. Equiprobable, Gaussian, or statistically independent variation is not required nor observed; blind does not mean uncaused, unintentioned, or unengineered (D.T. Campbell, 1974). Blind means lacking perfect foresight with respect to selection.

What are these sources of variation? In the organic world, mutation is a weak force of variation. Instead, introgression of traits from conspecific and heterospecific populations, including hybridization, is the strongest force of variation (Grant, 1986). And organic-styled engineering efforts, including, I'd argue, sexual selection, seek to influence the production of variants in non-random directions. Within the active language replicator model, hybridization, the crossing-over of different texts, is prevalent and the engine of significant innovation. Perhaps text reproduction is more like fungi than mammalian reproduction. Fungi have tens of thousands of different sexes and mating can occur between all of them (Metzenberg, 1990). Such large-scale introgression may be analogous to the sources and scale of text innovation. (See Pinker (1997) for a discussion of adaptive directional mutations in cultural evolution.)

Thus, I claim, that even in the presence of engineering or languagemaking on the part of individual authors, we do not have an explanation of the accumulation of complex design at the level of the language itself, since the engineering of variation (directed as it may be) does not cause retention and reuse trans-generationally and it is this that is the source for complex functional design. However, I still need to explain adaptive complexity of the language without calling on some socio-cultural *uber* engineer.

5.7.1 Nonselectionist forces

It has been argued that selection is the only known force, other than direct engineering, that can account for functional adaptive complexity (Pinker & Bloom, 1990; Cosmides, Tooby & Barkow, 1992; and elsewhere). However, to assume that because I have adaptive complexity without "invisible hand" engineering I, therefore, have selection, would beg the question I am here proposing (in the aristotelian sense). Instead, I will review, and rule out, the potential nonselectionist forces that have been advanced for organic evolution, and I will consider the forces related to transmission mechanisms that are distinct to language and socio-cultural evolution.

Pinker and Bloom (1990) have listed a collection of nonselectionist and nonadaptationist mechanisms that could conceivably account for (in particular, according to Gould) the accumulation, in an organic system, of a particular complex trait over time: "genetic drift, laws of growth and form (such as general allometric relations between brain and body size), direct induction of form by environmental forces such as water currents or gravity, the effects of accidents of history (which may trap organisms in local maxima in the adaptive landscape), and 'exaptation' (Gould & Vrba, 1982), whereby new uses are made of parts that were originally adapted to

some other function or of spandrels that had no function at all” (Pinker & Bloom, 1990, p. 709).

In order to establish selection as a strong force, I need to consider and rule out these other forces. First, I can easily rule out drift, because under conditions of random drift we do not expect the strong and frequent trait/fitness covariances between the replicators and their populations which I have observed (Endler, 1986). Furthermore, drift should only function within rather small population sizes.

Allometric dynamics are indeed observed within the chronica. Namely, the larger a text, the more traits possessed. Thus, I would expect that the larger a text the higher the level of expression for some particular linguistic replicator, relative to a shorter text. This is an allometric growth, of a sort, and could account for the accumulation of some traits if, in particular, we witnessed a steady growth in text length. To account for this dynamic, I always normalize all traits by the length of the documents. The metric traits associated with any collection of replicators express the percentage of text devoted to that replicator, and not the absolute number of occurrences. (This kind of normalization is not unlike that done with brain to body mass ratios used to establish rates of human encephalization (e.g., Deacon, 1997)).

Physical forces are certainly important to organic evolution. After all, thermodynamics has been around much longer than natural selection. What are the laws of physics within a collection of texts, or within the discussion systems of cyberspace? The particularities of each text environment define, it would seem, their own laws of physics. For example, within the Clinton chronicon a number of the lexico-syntactic replicators that occurred most frequently were due to a footer added by a popular NetNews discussion system (see Section 4.4.2):

```
-----== Posted via Deja News, The Discussion Network -----  
http://www.dejanews.com/          Search, Read, Discuss, or Start Your Own
```

The inclusion of this banner is, in a sense, part of the laws of physics within the DejaNews environment; friction produces heat, and DejaNews produces footers. Similarly, the lexeme “Re:” is added to in-reply-to posts by most NetNews systems. The cyberspace laws of physics do, then, account for some of the dynamics I have observed. However, I have not observed these features having evolutionary force, as revealed in a trait/fitness covariance. For instance, while elements of the DejaNews footer occur on the list of most frequent replicators within the Clinton chronicon, they do not appear as active replicators. Furthermore, these environmental forces

are generally easy to observe and account for. It would be hard to be fooled by them.

A nonselectionist accident that could account for a design as complex as the “Nazi” trait, for instance, would be quite an accident indeed. This would amount to authors accidentally using the term “Nazi” often enough and in more and more general and pejorative contexts so as to establish its socio-cultural currency in those contexts. Thinking that only chance can produce such complex design is like, as famously argued by Hoyle and Wickramasinghe (1981), a tornado blowing through a junkyard and assembling a 747 jet. In fact, the process of selection does exploit small bits of luck and chance in the accumulation of adaptation and complex design (Dawkins, 1983). But luck, on its own, can not come close to producing something as complex and functionally designed as this.

Finally, the flaws and fallacies around arguments of exaptation and spandrels have been so forcefully revealed by others (in particular Dennett, 1995) that I am comfortable with dismissing such nonselectionist “forces” by simply summoning the power of these arguments. There comes a time to move on.

In summary, I believe that chance, drift, exaptations, and spandrels do not figure as prominent forces (or are meaningless concepts to begin with). I do believe that laws of growth and the physics of cyberspace are a cause for some observed behavior. But, their contributions are easily accounted for and have not demonstrated significant evolutionary capacities.

Thus, I have accounted for the major nonselectionist forces considered within organic evolution. But are there nonselectionist forces that might be particular to language evolution or cultural evolution in general? Indeed, Cavalli-Sforza and Feldman (1981), Durham (1990, 1991), Boyd and Richardson (1985), Lumsden and Wilson (1981), and others have noted that transmission forces are potentially significant elements within cultural evolution. By transmission forces, they mean things such as the relationship between a teacher and learner, the generational differences and numerical relations between teachers and those taught, the complexity of a society, and so forth (Cavalli-Sforza & Feldman, 1981, p. 62). I would add that technologies can dramatically impact the various transmission rules, for instance, the advent of NetNews affords new modes of transmission that had previously been unavailable.

I am sure that transmission forces play a critical role in the dynamics of language replication and this remains an important area for further study. For instance, does

the quality of language evolution vary between texts on the net (e.g., the Clinton chronicon) and those that are print based (the Globe chronicon)?

It cannot be the case that transmission forces alone act as the sole source of adaptive complexity, since these forces (at least as listed above) are too invariant to account for the rapid variational dynamics I've observed. Cavalli-Sforza & Feldman (1981) argue that considerable progress can be made on theories of evolution dynamics without having developed a complete understanding of transmission mechanisms. That's why Darwin was able to make substantial progress without knowing what Mendel knew (let alone Watson and Crick).

5.7.2 Neutral models

In Section 3.11, I made use of neutral shadow models texts and replicators in order to establish that correlations between traits and population success were not due to simple structural properties of the text collections (e.g., word frequencies, text lengths, clustering methods). A number of random evolution models have been proposed in order to study (or rule out) nonselectionist and nonadaptationist features (Raup & Gould, 1974; Gould, Raup & Sepkoski, 1977; Bedau, Snyder, Brown & Packard; 1997). The central feature of a neutral model is that it removes any link between traits and individuals and the subsequent survival of individuals and retention of traits. As such, any accumulation of usage would necessarily be due to chance, historical accident, physical laws, or other nonselectionist forces. Figure 54 shows again the correlation coefficients from the neutral shadow model (with original clusters) first shown in Figure 20. This neutral model was created by randomly permuting the weights of all term vectors within the term/document matrix and then attempting to discover active replicators. Clearly, the random model finds no strong covariance, and thus one can argue that the significant trait/fitness correlations of

active replicators must be due to qualities of the traits themselves, in contrast to chance, accident, simple properties of structure or algorithms used, and so forth.

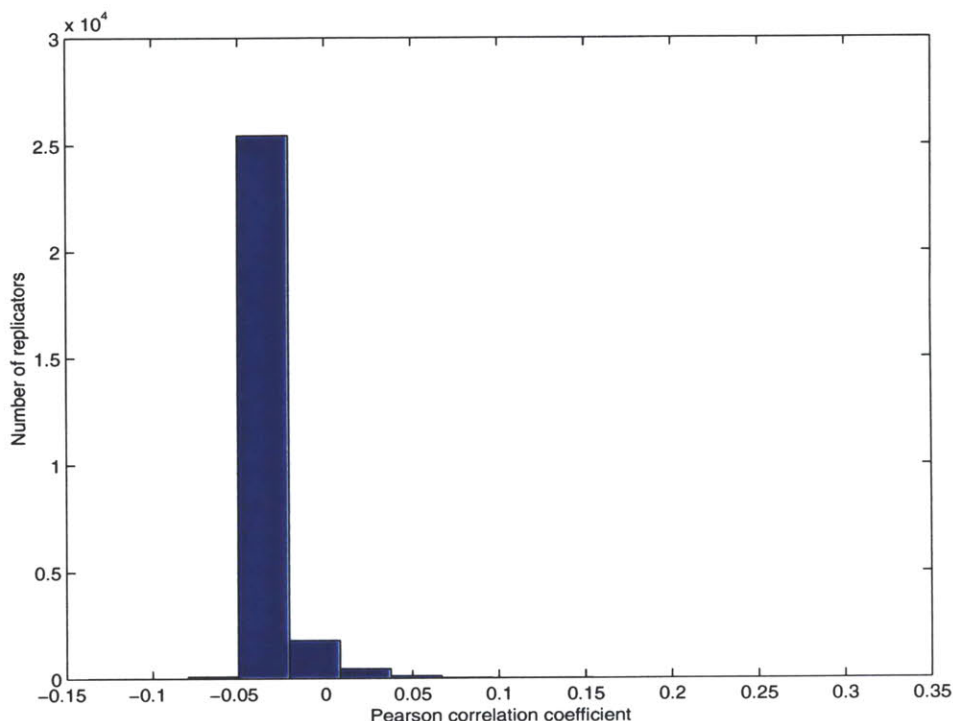


FIGURE 54. Histogram of correlation coefficient for all lexical replicators from neutral shadow model of Clinton/Lewinsky reporting within Globe chronicon (reproduced from Figure 20).

5.7.3 Accumulation of usage

I have been arguing that complex design within language is due to the biased survival of variant forms due to selection. This suggests that language elements undergoing active replication should accumulate greater and greater *usage* than would otherwise be expected since these forms are being selected. This accumulation of usage (e.g., a word occurs more and more frequently with time) should correlate with the accumulation of design (e.g., this word becomes more well fitted to the information processing environment). Indeed, it has been discovered that the accumulation of complexity, as measured by richness of meaning, correlates with the

accumulation of word usage; the number of occurrences of a word statistically relates to the number of different senses it realizes (Sinclair, 1991).

This suggests a simple question easily tested: Do active lexical replicators accumulate more usage over time compared to the population of words in general? That is to say, are active replicators accumulating more usage than would be expected by chance. Given the observation of Sinclair, this would suggest that these active lexical replicators are accumulating complex design as compared to the set of words at large.

I developed a system to measure the number of occurrences of each word over time across a chronicon. Phenomena of this sort is best measured over a fair period of time, so I studied the Globe collection which represents nearly two years of texts. I discretized time into roughly one week buckets and counted the number of times each word occurred in each bucket. This created a timeseries of usage for each word, t_0, t_1, \dots, t_k . It is the relative growth (or diminution) of usage that is of interest here. To determine this I simply subtracted the usage statistic at time t_n with the value at time t_{n-1} . This, then, describes a timeseries of the relative growth in usage for each word. And the average of each of these values summarizes the relative growth in usage for that word over the entire collection of texts. Table 32 shows the mean accumulation of usage and the standard deviation about the mean for the Clinton/Lewinsky and Iraq clusters within the Globe collection. I first show the values for all words and then those for the top active lexical replicators from Table 12 and Table 13.

Cluster	All words, mean	All words, standard deviation	Active replicators, mean	Active replicators, standard deviation
Clinton/Lewinsky	0.0029	0.0125	0.0488	0.0430
Iraq	0.0000	0.0026	0.0058	0.0076

TABLE 32. Accumulation of usage of all words and active lexical replicators from two clusters within the Globe collection.

Clearly, the active lexical replicators are accumulating usage more than would be expected from chance or due to, for instance, a gradual increase in volume of texts within a cluster. This is born out by statistical hypothesis testing. The null hypothesis, that the two sample means are equivalent, is rejected by the two-tailed t-test with a significance level, $p < 0.000001$.

Note that the active replicators from the Clinton/Lewinsky cluster are growing in usage on average by about 5% per week across the two years of texts, while the population of words in general grows by roughly 0.3% (which probably is due to a general increase in volume of texts over the time period).

5.7.4 Summary

I believe the only available conclusion is that active replicators are accumulating usage and design due to selection as a strong force. It is worth stating explicitly that selection here, while natural, is not “natural selection” in the customary sense. Instead, it is a form of *cultural selection*. Cavalli-Sforza and Feldman (1981, p. 15) define cultural selection as the rate or probability that a given innovation, skill, type, trait, or specific cultural activity or object will be accepted in a given time unit by an individual representative of the population. The selective process posited by the Microevolutionary Language Theory is slightly different. I imagine cultural selection acting on the individual texts, with heavy hybridization as a source of considerable cross-over between multiple individuals.

David Hull has noted that the processes of cultural selection might be intention and purpose driven, similar to the sources of variation; the cultural milieu might select based on “conscious agents doing things for a purpose” (Hull, 1999a). This is not inconsistent with the evolutionary algorithm in general, with organic evolution (Hull, 1982, 1999a), nor with language or socio-cultural evolution (Durham, 1991).

In summary, I have argued for a de-coupling of the individual production of variant forms and any socio-cultural selective forces. And, at the socio-cultural level, I have argued against potential nonselectionist forces and claimed that mechanisms of transmission, while critically important, can be somewhat independently studied and, at the least, should not impede us from making progress where we can.

5.8 The Size of the Units of Selection

In Chapter 4, I distilled active replicators at four linguistic levels: lexical, lexical co-occurrence, lexico-syntactic, and syntactic. But is one or some of these levels the true and precise target of selection? In other words, how do we know the right *size* for a replicator? In Section 5.5 I argued that all of these linguistic levels were suitable units of selection. But now I wish to determine if any one of them is more likely to be the precise target of selection (for which the other levels act as traces to, or portions of).

Following Williams' definition of the evolutionary gene as "that which segregates and recombines with appreciable frequency" (1966, p. 24), I argue that: The appropriate target of selection will be *the largest units of socially transmitted information that reliably and repeatedly withstand transmission* (Pocklington & Best, 1997).

This definition describes a unit that is most likely to come under selection and, thus, respond with the production of adaptations. While a gene is perhaps more appropriately defined as an open reading frame in the DNA, or a section of DNA that creates a single protein transcript (Watson, 1976), Williams' definition of an evolutionary gene still has utility. The two important aspects of this definition are that a unit be large enough to exhibit properties that can covary with replication success, and still be small enough to have robustly developing characteristics that reappear from host to host.

An unclear perspective on the precise locus of selection can confuse our understanding of evolution and cause us to waste time looking for adaptations where none are likely to exist. I've argued that the process of adaptation depends upon units of selection which possess variable properties that can be modified. As these units become smaller (lexemes, graphemes, etc.), they will provide less raw substrate on which selection can act. As units become larger (phrases, sentences, entire texts), they will fall prey to two problems, both of which will cause them to be less likely to generate adaptations. First, they will become less likely to reproduce with sufficient fidelity, due to the larger number of external contingencies involved in their replication process. Second, they will be subject to fewer sorting events. Sorting events are instances where one alternative versus another is differentially replicated.

Thus, larger units (presumably replicating less frequently) will be subject to selection as a weaker force (as they undergo fewer sorting events), as well as being ineffective at responding to selection when it does occur (due to their lower replicative integrity). The size of the units will represent a trade-off between increased substrate, on which selection can act, and the twin problems of reduced selection pressure (due to fewer sorting events) and reduced effective response to selection (due to contingencies). In any case, whatever the large units may be, they are composed of the smaller units and thus I assume some sort of hierarchical organization. Even if the most precise target of selection is larger than a word, for instance, words may demonstrate evolutionary significance. For a discussion and review of hierarchical organization schemes of cultural replicators and their parallels in biological systems, see Sereno (1991).

Are words simply traces to larger targets of selection, such as noun phrases? Are there even larger targets? And how do syntactic replicators fit into this?

Words are the smallest replicators I have examined. Some active lexical replicators clearly are components to larger targets of selection distilled by the CAMEL system. If we compare Table 11 with Table 20 in Chapter 4, it is clear that at least one lexical replicator occurs as a part of an active lexico-syntactic replicator: “Lewinsky” is part of “Monica S. Lewinsky.” But note that the “Lewinsky” replicator occurs more frequently than the entire noun phrase, “Monica S. Lewinsky.” Thus, the larger replicator does not wholly subsume all occurrences of the smaller one. Furthermore, note that every other strongly correlated active lexical replicator within the Clinton/Lewinsky cluster of the Globe chronicon does not appear as constituent to the larger phrasal replicators. A similar condition obtains for the lexical and lexico-syntactic replicators within other clusters. So words, in general, are not too small to be counted as legitimate targets of selection in their own right. If they were ignored, we would miss some replicators of evolutionary significance that do not appear as larger units.

Again, we wish for the *largest unit* which repeatedly withstands transmission. It is possible that there are replicators larger than words and lexico-syntactic units that are legitimate targets of selection. The clearest quantitative evidence, one way or the other, should come from the lexical co-occurrence analysis (see Section 4.3). The principal component method distills the largest co-occurrences reliably reoccurring across the chronica. Unhappily, though this seems like an ideal method, the results to date are not very promising. Most vectors returned from the analysis do not single out sets of words that can be readily determined and distilled; instead, it returns statistical smears across the entire term set. The results are inconclusive as to whether larger units of language will regularly act as targets of selection; phrases may, indeed, be the largest.

Finally, the syntactic replicators do not easily fit into the above hierarchy of size. However, one could explore larger and smaller syntactic units — from single parts of speech, to tuples, to larger n -grams. This is a legitimate research question: What is the largest syntactic replicator that acts as a target of selection?

Independent of whether one could find a larger syntactic unit that reliably and repeatably withstood transmission is the separate question: How much of syntactic traits are better captured by lexico-syntactic features? We are back to the question of whether it makes sense to attempt to untangle lexis and form to begin with. Note that this would not be inconsistent with studying single word replicators, since we could explore words with their part of speech as a 1-gram replicator.

In summary, determining the appropriate size (or sizes) for the targets of language selection remains an open research area, however, I have made progress in both framing the question and establishing bounds to an answer.

5.9 Macroevolutionary Consequences of Microevolution

Microevolution is to adaptive complexity as macroevolution is to speciation and diversity (Mayr, 1991). And if microevolution sits as the central problem within the evolutionary sciences (see Section 5.6), macroevolution runs a close second. Darwin did not think that these two processes were separate. He supported a direct link that has been referred to as “the macroevolutionary consequences of microevolution” (Plotkin, 1994, p. 59). This linkage was put on a solid footing by the work of George Gaylord Simpson (1944) who integrated macroevolution *qua* paleobiology with the neo-Darwinian synthesis (Eldredge, 1989). However, in recent years the sturdy relationship between large scale patterns of evolution and the microevolutionary processes of adaptive change has been challenged (e.g., Gould & Eldridge, 1977; Gould, 1980; S. Wright, 1982; but cf. Dennett, 1995).

Current controversies within organic evolution notwithstanding, the prospects for a strong link between the Microevolutionary Language Theory and language macroevolution is certainly seductive. By language macroevolution I mean, in particular, the glossogenetic (de Grolier, 1983; Hurford, 1991) program of historical linguistics which studies the evolutionary differentiation of language, e.g., Old English to Middle English to Modern English. That is, the history through which English became isolating, fixed-word-order, accusative, subject-prominent, SVO, etc. from originally being free-word-order, highly inflected, topic-prominent, etc. (Pinker, 1994, p. 232), along with the codependent history of the emergence of meaning and categories of meaning. A strong link between the Microevolutionary Language Theory and language macroevolution would establish that the sort of adaptive variation in traits I have arguably observed is sufficient, at some level, to describe the large scale differentiation of language. Clearly an exciting prospect!

Francisco J. Ayala (1983) has identified three subquestions as components to the larger question — does microevolution account for macroevolution? Here I list them as applied to language evolution: (1) Do the known microevolutionary processes operate on the individuals and populations that make up languages or language groups, and have they operated throughout the history of language; in other words, are the processes vertically and horizontally *pervasive*? (2) Are the microevolutionary processes *sufficient* to account for the large scale changes of macroevolution, or do additional (microevolutionary) mechanisms need to be proposed?

(3) Can the large scale trends of language macroevolution be deduced, are they *derivable*, from our understanding of microevolutionary processes? Pervasive, sufficient, and derivable: I will consider each of these issues within the context of language.

First, let's consider if the known microevolutionary forces are vertically (temporally) pervasive. Steven Pinker has noted that while there are Stone Age societies, there are no Stone Age languages (Pinker, 1994, p. 27). The various tongues, no matter how or when they branched from a common protolanguage (Hildebrand-Nilshon, 1995), share a striking similarity in formal complexity. Pinker was using this point to establish the innateness of language. But this aspect of language suggests (and it could well be true *ex hypothesi* an innateness theory) that any microevolutionary forces for adaptive complexity today (e.g., variation, drift, selection) are the same forces of yesterday, since formally languages of today are the same as languages of yesterday. The "physics" at play within linguistic systems has not changed from the origin of language differentiation. (This does not require that the evolution of and for language was frozen in one isolating moment in the Pleistocene, as seems to be suggested by some language instinct theories. Bates and MacWhinney have humorously addressed this essentialist-evolutionist debate (Messer, 1995) by comparing the essentialist program to Mario Cuomo's characterization of the anti-abortionist community who "believe that life begins at conception and ends at birth" (Bates & MacWhinney, 1990, p. 728).)

If these forces have been consistently present vertically through time, they are consistently present — for exactly the same reason — horizontally across language and language groups. The conclusion, then, is that the observed microevolutionary forces are pervasive across language and were as active in extinct languages as they are in extant ones.

It is another question whether these forces, in particular variation and selection, are sufficient to account for macroevolutionary differentiation; perhaps other forces, so far unobserved, need to be postulated. Establishing the *sufficiency*, in contrast to the *necessity*, of these microevolutionary processes is at least a tractable problem. An adequate review of the known macroevolutionary dynamics (for instance the genetic classification of language over time (Ruhlen, 1992)) might establish that all observed large scale patterns are obtainable from the known microevolutionary processes. However, such a complete review is beyond the scope of this dissertation.

It is, at the least, instructive to consider the recent controversies within organic evolution as they impact the sufficiency question. The outcome of this debate, in my opinion, has reaffirmed the sufficiency of the known microevolutionary processes

to account for macroevolutionary patterns. At the center of this controversy was the observation by Stephen J. Gould (Gould & Eldredge, 1977; Gould, 1980), not novel to him but infused with a new revolutionary drama, that the natural morphologies frozen in the fossil record show periods of considerable stasis punctuated by geological moments of great change. While this observation is true, the fact is that paleobiological data not only show this sort of “punctuated equilibrium” but contain the whole range of tempo and mode in diversity and innovation. As Eldredge put it, “many studies reveal the entire spectrum of possibilities within the same data set: while most characters are stable, some show progressive change, while others display change concentrated in relatively brief episodes, interspersed with vastly longer periods of no change at all” (Eldredge, 1989, p. 67).

The question which Gould proposed to answer in the negative was, could the known microevolutionary processes account for this range of temporal patternings or do additional mechanisms need to be operating (at the micro- or perhaps solely at the macroevolutionary scale)? It is now clear that the microevolutionary processes (primarily of variation and selection) are indeed sufficient to account for this range of large scale patterns. This has been shown most convincingly by a number of mathematical existence proofs, via formal modeling and computer simulations. Here, formal systems have been infused with only microevolutionary prowess and, nonetheless, express this range of macroevolutionary dynamics (e.g., Vose & Liepins, 1991; Green, 1993; Bedau, 1995). This formal work has resulted in a model of “self organized criticality,” proposed by Per Bak and co-authors (e.g., Bak & Sneppen, 1993; Flyvbjerg, Bak, Jensen & Sneppen, 1995). Here, a local system, obeying simple rules, without any global controls, and invariant to scale organizes into a critical steady state with occasional avalanches of all sizes. (The poster-child, self-organized critical system is a simple pile of sand!)

Finally, I turn to the question of derivability. Can the large scale patterns of macroevolution be derived from our understanding of the microevolutionary processes? Ayala (1983) demonstrates that this is not possible within organic evolution, at least in practice. In particular, the microevolutionary forces provide no way of predicting ahead of time whether a given macroevolutionary character will exhibit smooth transitions, over geological time, or catastrophic, punctuated ones. Thus, the choice between those competing patterns cannot be made, based only on knowledge of the micro-forces.

Moreover, even in theory, it is, in general, not possible to move from micro-deterministic forces to a macro-order (D. Campbell, 1989; Forrest, 1990; Cariani, 1991). This, thanks to the emergent outcomes of nonlinear and epistatic micro-processes being sensitive, for example, at arbitrary scales to initial conditions.

Thus, I believe that the microevolutionary forces within language are pervasive and sufficient to account for the macroevolutionary patterns of historical linguistics. However, these macro-dynamics may not be derivable directly from the microevolutionary forces. That notwithstanding, a considerable amount of macroevolutionary dynamics can nonetheless be framed, constrained, and elucidated by a thorough understanding of the microevolutionary forces.

5.10 Summary

The chapter began with a review of the two major models of contemporary evolutionary theory: the Lewontin-Campbell computational theory and the Dawkins-Hull typological theory. I used these to help frame the active language replicator model and to make progress on a variety of theoretical problems. This suggests a conceptual integration of contemporary evolutionary theory with corpuslinguistic models of language use.

By offering a demonstration of the Lewontin-Campbell algorithm within natural language, I gave substantial support for Campbell's Rule within text. That is to say, I demonstrated a Microevolutionary Language Corollary — organic evolution and language evolution are isonyms of one another. Next, I used the Dawkins-Hull ontology to examine the replicator as central beneficiary to the evolutionary process. Thus, the answer to *Cui bono?* within human natural language is the active replicator: a word, a phrase, and so forth.

The Microevolutionary Language Theory states that complex functional traits accumulate at the simplest level within language due to the evolutionary algorithm. In Section 5.6 I finally was able to demonstrate the development of a complex functional trait in language. "Nazi," as a pejorative attack word, has significance in a variety of environments and was selected amongst semantic variants due to some design quality it possessed.

I ended this chapter by surveying three important areas within current evolutionary theory: selectionist forces, the size of the target of selection, and the macroevolutionary consequences of microevolution. My arguments here are in the style of what Sir Karl Popper (1965) calls "bold conjecture," that is, reasoning which admits to tests against evidence and that welcomes attempts at refutation.

I have used the CAMEL software system to distill active replicators from within natural language. In the previous chapter, I framed these empirical studies under the dominant models of contemporary evolutionary theory. In this chapter, and the next, I will explore what can be *done* with this active replicator model of language.

The current chapter describes a set of experiments on ecological interactions between clusters of texts. I use a special, small collection of NetNews posts to show that some text clusters are in competition with others for the limited resources of authors and air-time on the net. The clusters which are in relatively narrow ecological niches within the information-processing environment, those that contain a smaller number of threads of discussion, are more likely to be in competition. This is similar to what is found within natural ecologies.

These results are examples of the sort of interaction studies that are suggested by the Microevolutionary Language Theory. Results of this sort can help to build further evidence for the theory. In the next chapter, I will show how active replicators can actually aid in a practical engineering problem, namely, text retrieval.

6.1 Models for Interacting Populations

The Microevolutionary Language Theory is primarily concerned with replicator dynamics and the accumulation of design at that level. However, having a model for

such processes, it is possible to explore richer population-level dynamics. For instance, can the Microevolutionary Language Theory and active replicator model offer any insights into *ecological* dynamics? Ecology here means a population of individuals interacting around some (generally scarce) environmental resource.

I have examined the pairwise interactions between clusters of texts within a chronicon. Pairwise interactions within populations have been widely studied within theoretical ecology. Consider two interacting populations: one population can have a positive effect on another by increasing the other's chance for survival and reproduction (+); or a negative effect, by decreasing the other population's survival chances (-); or a neutral effect (0). The ecological community has assigned terms to the most prevalent forms of pairwise interaction, in particular:

Mutualism (+, +)

Competition (-, -)

Neutralism (0, 0)

Predator/prey (+, -)

(Pielou, 1969; May, 1981).

My goal is to study the pairwise interactions of text clusters within chronica with the hope of discovering and better understanding some of these interaction types within natural language.

6.2 Special Test Chronicon

For this experiment, I employed a special (and smaller) chronicon. The collection consists of all texts posted to the soc.women NetNews newsgroup between January 8, 1997 and January 28, 1997. The soc.women newsgroup deals with a wide range of issues of interest to women. The chronicon consisted of 1,793 texts over this ten day period. The clustering mechanism arrived at 292 lineages the largest of which contained 103 texts.

6.3 Timeseries Cross-correlation

The first step to study the interactions between cluster populations is to return to the population timeseries used in other analyses. In Chapter 2, I first introduced these

timeseries; for instance, Figure 7 shows the volume of texts published over time to the Clinton/Lewinsky cluster within the Globe chronicon. For each of the 292 text clusters within the soc.women chronicon, I computed a timeseries. As usual, when computing this series, the texts were bucketed; for this collection, a bucket size of 24 hours was used.

To study the relationship between the timeseries of two cluster populations of text, I employ the cross-correlation function. The use of the cross-correlation to study bivariate processes, and timeseries in particular, is well known (Chatfield, 1989).

Each timeseries is normalized to be of zero mean and unit standard deviation; that is, I subtract off the mean and divide by the standard deviation. In this way, the cross-correlations will not be dominated by the absolute volume of text activity within some cluster, and instead, will be sensitive to both large and small-sized clusters. Assuming a familiarity with the regular covariance and correlation functions, the cross-correlation for two time series, X and Y , is given by

$$\rho_{xy} = \frac{\gamma_{xy}}{\sqrt{\gamma_{xx}\gamma_{yy}}}.$$

Here, $\gamma_{xy} = \text{Cov}(X, Y)$ and γ_{xx} and γ_{yy} are the variance of X and Y , respectively. Note this formulation only considers the cross-correlation for a zero time lag. That is, it considers how the two timeseries are correlated at identically matching points in time. With a nonzero lag, the cross-correlation would study cases when the two series might have correlations offset by some fixed amount of time. Since the time data is grouped into appropriate chunks (in this case 24 hours), the zero-lag cross-correlation will be sensitive to covariances which have a time offset as large as these bucket sizes; this builds into the timeseries an adequate time lag.

When the cross-correlation between two sets of data is significantly different than zero the two sets of data are in some relationship. A positive value means an increase in one series is likely to co-occur with an increase in the other series. A

negative value means an increase in one series is likely to co-occur with a decrease in the other series.

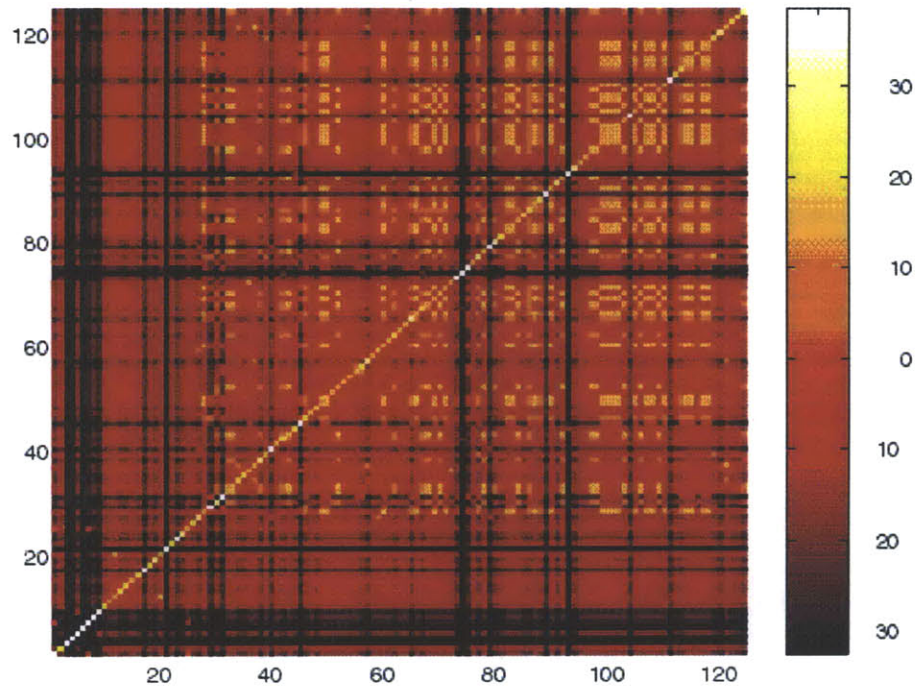


FIGURE 55. The pairwise timeseries cross-correlation for 125 largest text clusters within the soc.women chronicon.

Figure 55 shows the pairwise cross-correlations for the 125 largest text clusters within the soc.women chronicon. The diagonal represents the cross-correlation between a timeseries and itself which, as expected, is identically one. Note that the matrix is symmetric about the diagonal. The off-diagonal values range from near one to -0.26. The mean cross-correlation is 0.3. This value is high, indicating that many of these post clusters are positively related. I suspect this high average cross-correlation is, at least, partially due to external or systemic affects which were not removed by the bucket size. For instance, the analysis would be sensitive to patterns caused by the Monday-Friday work week common in the U.S. Further, some of this correlation may be due to a high level of mutualistic interactions amongst the texts

posted to the NetNews newsgroups. Clearly, the ideas conveyed within the soc.women newsgroup often share similar contexts.

In the analysis, this overall high correlation does not particularly matter; what really matters is the relative cross-correlation — that is, those that are the largest and those that are the smallest.

6.4 Negative Cross-correlations: Competition versus Predator/Prey.

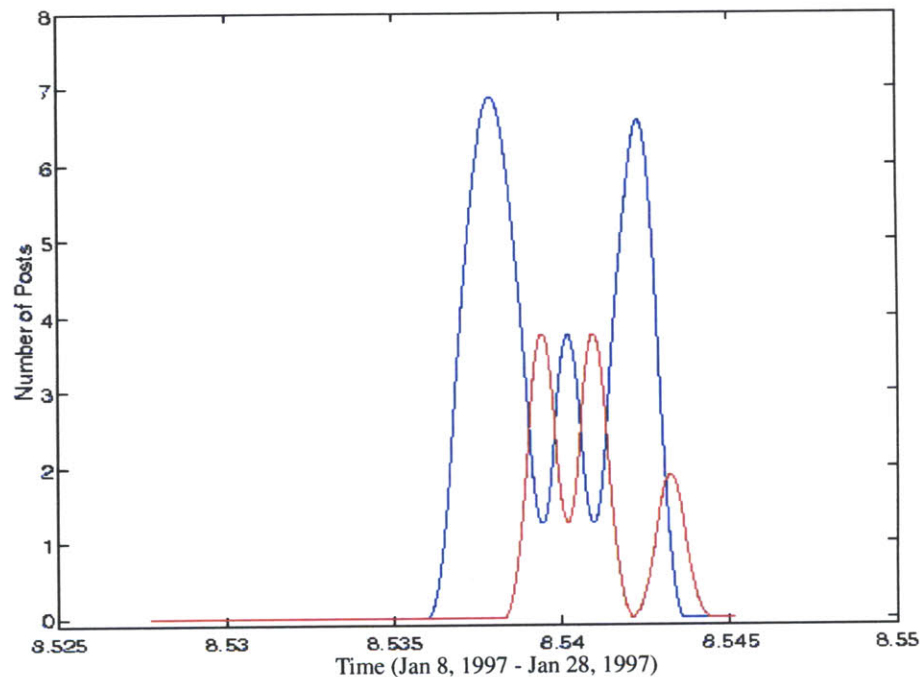


FIGURE 56. Volume of activity for two clusters. The cross-correlation between these two timeseries is -0.26

The pairs of text clusters primarily studied are those with relatively strong negative cross-correlations; to wit, those where $\rho_{xy} \leq -0.2$. Note that in all such cases (there are 42) $P < 0.001$, suggesting that with high probability the correlations are not due

to chance. Figure 56 and Figure 57 plot two such interactions, both fairly characteristic of this population.

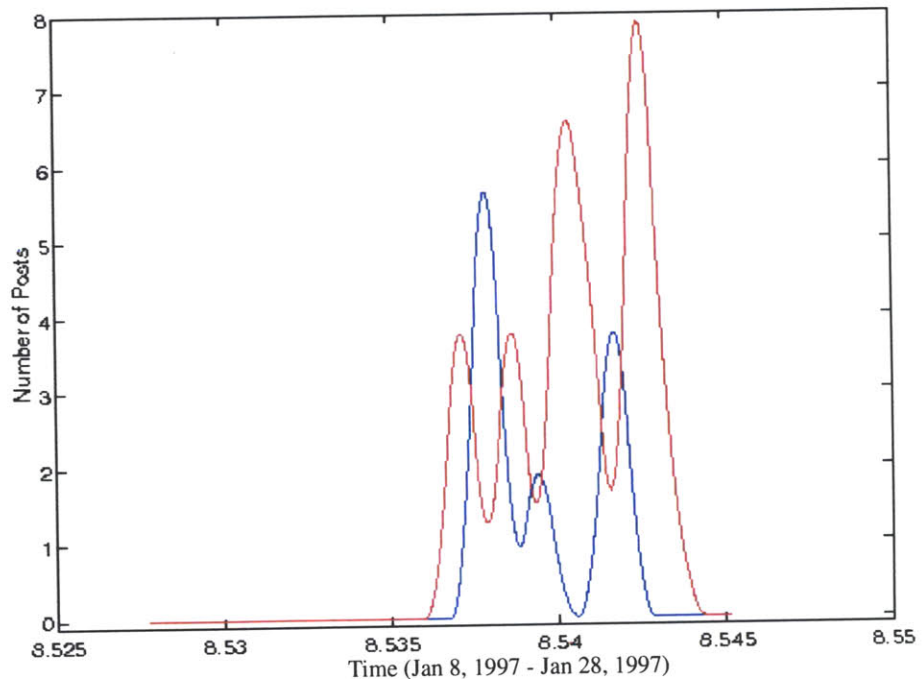


FIGURE 57. Volume of activity for a different set of two clusters. The cross-correlation between these two timeseries is -0.23

Both of these figures demonstrate a clear negative covariance between the volume of activity of the two clusters. This negative covariance is both statistically significant and visually compelling. But what do these graphs signify? Can this be interpreted within the rubric of ecological interactions?

At first glance, the interactions appear to be of a predator/prey variety; they have a (+, -) relationship to them. However, competition might also produce similar interaction phenomena, if the competitors are operating close to some limitation or environmental carrying capacity. In such instances, the relationship between population sizes will be a zero-sum game: when one goes up, the other must come down. To be able to classify the interactions of Figure 56 and Figure 57, I need to consider the qualitative details of these two interactions through direct study of the texts.

Recall that in the case of a predator/prey relationship, one population enjoys an increased growth rate at the expense of another population (e.g., one population feeds on the other). The presence of a relatively large population of predators will result in a diminished level of success for the prey (they get eaten up). Conversely, the relative absence of prey will result in diminished success for the predator (they have nothing to eat).

Now, consider the case of competition. In competition, two interacting populations inhibit each other in some way, reducing each other's level of success. This often occurs when the two populations rely on the same limited resource. Unlike the predator/prey relationship where the predator requires the prey for success, with competition, the two populations would just as soon avoid each other altogether.

This pressure towards avoidance is the source of much ecological diversity since it propels populations to explore new and, therefore, competition-free niches (Pianka, 1981). An ecological niche, for some particular species, is simply that collection of resources the species relies on. Interspecific niche overlap occurs when two or more species share one, some, or perhaps all of their resources. When those resources are scarce, interspecific competition will result. The width of a niche is simply an accounting of the variety and number of resources a population makes use of.

6.5 Competition and Niche Behavior

The texts that make up the four clusters represented in Figure 56 and Figure 57 have been studied closely in an attempt to classify their interactions. The two clusters of Figure 56 are both made up of posts within a single in-reply-to thread. The subject line for these posted texts reads, "Men's Reproductive Rights." In general, these posts are concerned with the responsibilities and rights of men towards their unborn children. The cluster displayed with a dashed line in the figure is centered around the use of contraceptives. It consists of a collection of texts wherein the authors debate who is most responsible, the women or the man, for using contraception. The clusters with a solid line deal with the use of abortion, and whether the father has any intrinsic rights in deciding to abort or not to abort an unborn child.

In Figure 57, the two clusters are also from a single in-reply-to thread. The subject line here reads, "Unequal distribution of wealth?". This particular discussion thread was rather large. There was a total of 365 texts posted to this thread, which the CAMEL software system broke into a number of clusters, due to significant bifurcations of topic. In other words, many parallel discussions occurred within a single in-reply-to thread. The cluster of texts shown with the solid line in Figure 57 cen-

tered around a debate whether the US military was a “socialist collective.” The cluster with the dashed line was a debate on the value of releasing the mentally ill from hospitals. Clearly, these two debates are quite dissimilar, even though they span the same set of days and are posts to the same discussion thread.

The two clusters of Figure 56 are different, but related, discussions. Those of Figure 57 are different, and not clearly related. Still, I believe that both of these sets of interactions demonstrate elements of competition. Within the texts there is no evidence of predation; in Figure 57, the topics seem entirely orthogonal to one another. However, in both examples, these texts (and their constituent replicators) are competing for the same collection of human authors who must act as agents, if they are to propagate and succeed. This seems even more likely, when we consider that all these posts are to the same newsgroup which, due to its narrow subject area, supports only a limited supply of human posters. Moreover, each pair of interactions is confined to a single thread of discussion, which, again, has an even more limited set of potential human authors, since users of the NetNews system often zero-in on particular threads they find interesting and ignore others. After inspecting most of the interactions which demonstrated strong negative correlations, I observed no examples of predator/prey interactions, but many instances which appeared to be examples of competition.

6.6 Competition

I have argued that these interactions are of a competitive nature; now, I'd like to test that theory. Again, recall that competition is often caused by populations existing within the same (narrow) ecological niche. What makes up an ecological niche for a text within NetNews? I propose to model the newsgroups themselves as spatially distributed ecological niches. Since there is relatively little interaction between newsgroups (save the phenomenon of cross-posting), one would expect these niches to behave something like island ecologies — they remain relatively isolated from each other. However, within a *single* newsgroup (e.g., soc.women), niches might be described by threads of discussions. As previously stated, I have found that individual posters to the system tend to become involved in in-reply-to threads which interest them. Thus, the texts within a particular thread interact with a set of human resources which is smaller than the entire set of potential human resources available to the newsgroup. These resources define niches within a newsgroup.

My hypothesis is that the cross-correlations which approach -1 in the chronicon seem to be examples of competition, and competition will be more likely between populations which are posted to the same threads and, thus, have overlapping

niches. The most direct way to test this theory is to see if negative cross-correlations between two clusters correlate with the degree to which they post to the same threads. For each of the 125x125 pairwise interactions, I computed the number of threads each pair of text clusters had in common, and divided that by the total number of threads posted to each cluster. For example, one cluster of texts may contain posts which went to two different in-reply-to threads. Another cluster may have posts which span three different threads, one of which is identical to a thread within the first group. So this pair of clusters would have posted to a total of four different groups, one of which was shared. Their relative niche overlap would, therefore, be 0.25.

I calculated the correlation coefficient between the negative cross-correlations of Figure 56 and the percentage of thread overlap between these pairs of clusters. I found this correlation to be -0.04. While this correlation is statistically significant ($P < 0.001$), it is not very pronounced. The negative sign, though, does indicate that as the level of competition increases (a negative cross-correlation) the percent of overlap of their niche also increases (a larger, positive, shared thread percentage).

This small correlation coefficient may be due to a small signal/noise ratio. Since most pairwise interactions result in small cross-correlations, the relative number of large negative correlations is quite small. The number of interactions grows with the square of the number of clusters. I suspect that a simpler experiment which grows linearly with the number of clusters will have a better signal/noise ratio.

I have also studied the correlations between the absolute number of in-reply-to threads of a cluster and the average degree to which the cluster finds itself correlated with all other clusters. My hypothesis is that the absolute number of threads a cluster is posted to will be related to the average degree of competition the cluster experiences in its interactions. Since the variety of resources used by an entity defines its niche, if a cluster of texts is posted to a relatively small number of threads, it exists in a narrow ecological niche. Should there subsequently be any interspecific overlap of these narrow niches, scarcity will result in competitive encounters. I computed the correlation coefficient between the total number of threads within a cluster and its average cross-correlation value. The correlation coefficient here is 0.25. Thus, as the number of threads within a cluster increases (the set of available resources is widened) the average level of competition diminishes (the mean pairwise cross-correlation also increases). This correlation is statistically significant ($P < 0.001$) and more pronounced.

I also computed the correlation coefficient when the absolute number of threads was normalized by the size of the clusters. One might expect that the number of

threads employed by a cluster would grow with the number of texts posted within that cluster; in other words, as a cluster gets larger, the number of threads increases too. This might affect the analysis above, such that, instead of measuring niche width, I was simply measuring cluster size. Dividing out the size amounts to computing the average number of texts per thread for a given cluster. When this set of values was correlated with the mean cross-correlation, I arrived at a nearly identical coefficient as above, which again, had a clear statistical significance. Thus, cluster size is not a major factor in level of competition.

6.7 Summary

This chapter offers an example of a sort of study made possible by the active language replicator model. In particular, I examined the pairwise interaction between clusters of posts to soc.women by computing the cross-correlations between their timeseries. For cases of strong negative cross-correlations, I theorized that this may signify conditions of competition between the interacting populations where the clusters are competing for a limited set of human authors. As support for this theory, clusters with relatively narrow ecological niches, those which make use of a small number of in-reply-to threads, are more likely to be in competition with other clusters of texts. This behavior is analogous to what is found in natural ecologies (Pianka, 1981).

Why do these clusters compete? Qualitative analysis of the posts, such as those described in the previous section, shows that many competing clusters are composed of posts sent to the same or similar threads. Competition is over the scarce authorship resources within these specific thread niches. My speculation is that, over time, a particular thread of discussion may divide into two, or more, internal themes which then proceed to compete for “air-time” within the thread.

Replicators and Text Retrieval

In the previous chapter I showed how active language replicators and the Microevolutionary Language Theory can frame a study of ecological interactions between populations of texts. In doing so, I demonstrated the sort of research programs that are supported by this work. But the best theories, in my opinion, should also suggest practical results — solutions to real and tangible problems.

In this chapter, I will apply the active language replicator model to a very practical and timely problem, namely, text search and retrieval. With the growth of the World Wide Web, and other online text collections, retrieval is becoming a very significant engineering problem. How can you find texts that are of interest to you amongst the millions of irrelevant documents? A number of researchers have attempted to use natural language processing techniques, such as the inclusion of phrase information, to aid retrieval tasks. I have found that by using active lexico-syntactic replicators as features in a text search, I am able to materially improve retrieval.

7.1 Review of Text Retrieval Experiments

In Chapter 3, I overviewed the core methods used by text retrieval engines and shared with the CAMEL system. In Section 3.8, I reviewed the vector space representation wherein texts are scored on the relative presence or absence of terms found within the collection. This process assigns a vector to each document in the collection; this vector places the document within a high-dimensioned (on the order

of 10,000, usually) space of words. Then, in Section 3.9, I reviewed the cosine similarity metric. This function measures the distance between the vector representations of two documents (or, as we will see, between a document and a query).

Given a collection of texts, each represented by a point in the term vector space, and the cosine similarity measure, we have all that is required for a basic text retrieval engine. Thus, the CAMEL system, without any real modifications, is already a text retrieval engine.

Consider the Globe's collection of 22,498 texts. A retrieval task might pit a query, such as, "give me all documents dealing with the AIDS virus," against this collection of texts. The basic retrieval engine measures the "distance" between the terms of the query and each document's vector representation. The texts that are closest to the query are returned to the user as a document set. This retrieval method, that employs only term-frequency information in isolation, is sometimes referred to as a "bag of words" approach.

The National Institute of Standards and Technology of the U.S. Department of Commerce has sponsored, for the last seven years, an important annual conference on text retrieval (e.g., Harman & Voorhees, 1994). Central to these meetings, entitled the Text REtrieval Conference (or TREC), has been a series of competitions in which researchers pit their search engines against one another in a variety of retrieval tasks proffered by the TREC organizers.

These competitions are ruled by a standard methodology for evaluating the results from each search engine. The most common of these evaluation techniques is to measure the *precision* of the retrieval as a function of *recall* (Salton & McGill, 1983; Frakes, 1992a). Recall is defined as the ratio of relevant documents returned by the search engine to the total number of relevant documents within the collection. Consider, for example, our AIDS query against the Globe document set. A set of human judges determined that there are exactly 137 texts within this chronicon that are relevant to this query; in other words, the ideal search engine would, given this query, retrieve these 137 texts, and no others, from the chronicon. The recall, then, for some particular search is the percentage of these 137 texts returned. If the delivered document set contains 50 of these 137 texts then the recall rate is

$\frac{50}{137} = 0.37$. Precision, in contrast, is the ratio of documents returned that have

been judged as relevant to the total number of documents returned. If the search engine returned 50 relevant documents for our given query, but also returned 25 documents not judged relevant, then two-thirds of the documents returned are

judged relevant, and the precision is $\frac{50}{(50 + 25)} = 0.67$. Note that both precision and recall fall on the closed interval from 0 to 1.

In practice, the higher the recall the lower the precision: As a search engine returns more and more relevant documents, it becomes more and more likely it will return irrelevant documents as well.

Given these two measures for a retrieval task, a common way to evaluate a search engine is to plot its 11-point precision/recall graph for a set of queries over a particular corpus. In Figure 58 I show a fictitious 11-point precision/recall graph. Plotted on the graph is the precision of a retrieval operation given eleven fixed values for recall (0.0, 0.1, ..., 1.0). Let the total number of relevant documents for this particular query be represented by χ . Then, the number of relevant documents that are required for the particular recall value shown on the 11-point graph is given by

$$f(\text{recall}) = \text{recall} \left\lfloor \frac{\chi}{10} \right\rfloor + \left\lfloor \frac{\chi}{10} \right\rfloor,$$

for the values, $\text{recall} = 0.0, 0.1, 0.2, \dots, 1.0$. Thus, given our example query with

$\chi=137$ relevant texts, $f(0.0) = 7$, $f(0.1) = 20$, ..., $f(0.9) = 124$, $f(1.0) = 137$.

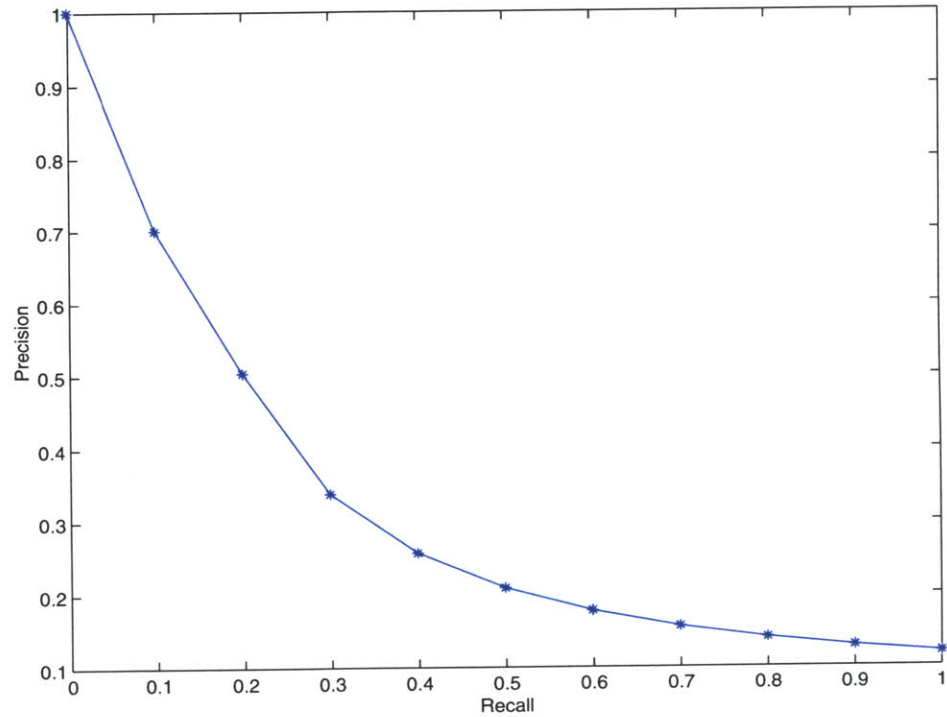


FIGURE 58. Fictitious 11-point precision/recall graph.

7.2 Queries

In this retrieval experiment I have evaluated the CAMEL search engine, using the 11-point precision/recall method, on twelve different queries against the Globe chronicon. I compared the results from using only the bag of words method against results obtained when term frequencies are augmented with active noun phrase replicators.

As mentioned above, the TREC conference has sponsored annual competitions in text retrieval. For these competitions, they have assembled a set of general *topics* from which researchers are free to construct queries. The motivation for using more general topics is to allow the individual researchers some leeway in constructing the actual queries (Voorhees & Harman, 1997). The TREC ad hoc task is one of the

most important tests of the competition. In it, the document collection is known ahead of time by the various researchers, but the queries (or topics) are not.

In Figure 59, I show an example AIDS query composed in TREC-3 ad hoc long-format. This is the format that I have employed for my current experiment. (See Voorhees & Harman (1997) for a review of the various TREC tests and formats.) The long-format is admittedly verbose. Recent TREC conferences have adopted shorter formats in order to better match the sort of queries normally composed for Web search engines.

<title> The AIDS virus

<desc> Relevant documents will discuss the AIDS virus, HIV. Documents could mention modes of viral transmission, such as unprotected sexual contact or needle sharing among intravenous drug users. Documents could also mention treatments for the infection including AZT, protease inhibitors, and combination drug therapy. Mention could also be made of medical research aimed at treating the illness.

FIGURE 59. Example AIDS query in TREC-3 ad hoc long format.

All of the twelve queries used in the current experiment are shown in Appendix A. These queries were developed by two students working on my behalf, but without knowledge of the intended retrieval experiments. These same students hand coded all 22,498 texts of the Globe chronicon against these twelve queries, marking texts as relevant or not to each of the queries. This coding took approximately 400 hours. (Note that, for the TREC conference, relevance evaluations are made by employing a less labor intensive polling method (Sparck Jones & van Rijsbergen, 1975).) In Table 33, I gloss each of these queries, and give the total number of documents

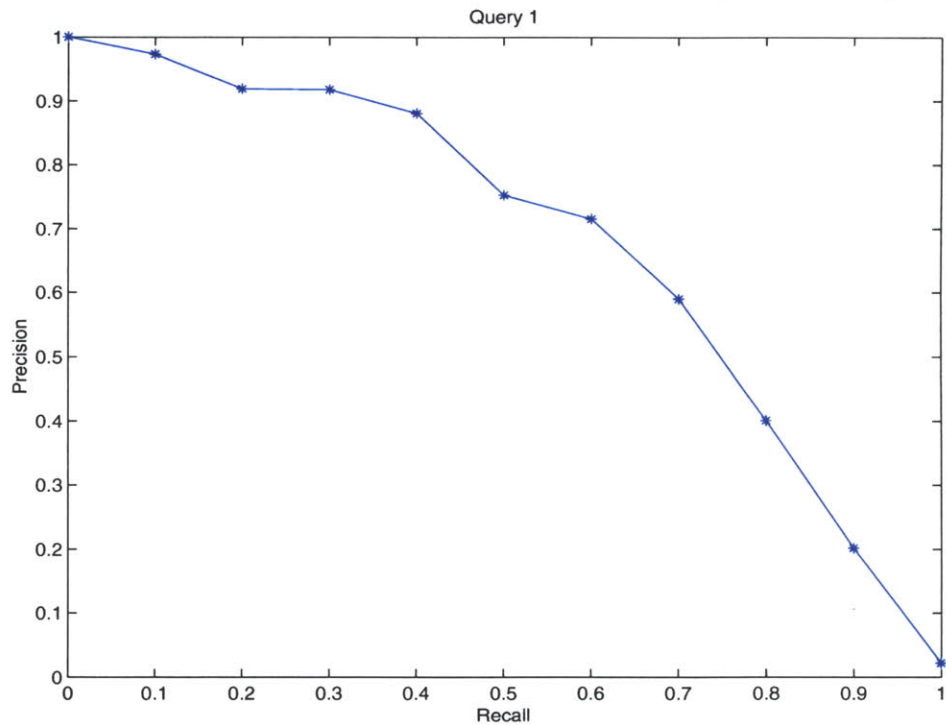
from the Globe chronicon that were determined to be relevant to each query by the judges.

Query	Query gloss	Number of relevant texts
1	Airplane crashes	324
2	Clinton campaign finance controversies	199
3	Congress reacts to immigration	36
4	AIDS virus	137
5	Sexual scandals within US military	126
6	Oklahoma City Federal Building bombing	151
7	Tobacco lawsuits	194
8	Genetic research	73
9	Cancer	139
10	Sexual scandals of Bill Clinton	605
11	Space exploration	294
12	Conflict in former Yugoslavia	371

TABLE 33. Gloss of twelve queries used in retrieval experiment, and number of texts judged relevant from the Globe chronicon.

In Figure 60, I show the complete long-format for the first query and the 11-point precision/recall graph that resulted from using the CAMEL system on this retrieval task. For this retrieval, only term frequency information was employed. The

CAMEL software system performs reasonably well against this query; this particular precision/recall graph is of fairly standard form, and the performance is good.



<title> Topic: Airplane Crashes and other Aviation Accidents

<desc> Narrative: A relevant document will refer to a plane crash, airliner crash, airplane crash or other aviation accident or air disaster. It may discuss investigations of the cause of the accident conducted by the Federal Aviation Administration or the National Transportation Safety Board. These investigations may concern the aviation industry as a whole or may deal with elements of the planes such as the flight recorder or voice recorder.

FIGURE 60. TREC-3 long-format and 11-point precision/recall graph for Query 1. These retrieval results are from the CAMEL system, using only the bag of words (term frequency) approach.

I want to compare the results from using only term frequency information with the results obtained when active noun phrases are added. To do so, I will employ a quantitative method to compare between 11-point precision/recall graphs. An improvement of one approach over the other can be measured as the percentage in increased precision averaged over all eleven points of recall. For example, the mean precision across the eleven points in Figure 60 is 0.677. Should some other retrieval experiment admit to an average precision of 0.80 across these eleven points then that would be an increased precision of 18%, since

$$\frac{(0.80 - 0.677)}{0.677} = 0.18.$$

Karen Spark Jones (Spark Jones & Bates, 1977) has argued that we should classify improvements above 5.0% as “noticeable” and those above 10.0% as “material.”

7.3 Natural Language Text Retrieval

The application of rich linguistic features, such as those arrived at through natural language processing techniques, to the problems of text retrieval has had a long history. Natural language enhanced retrieval engines have employed part-of-speech tagging, light syntactic parsing, phrase extraction, and related approaches to enhance the traditional bag of words approaches (e.g., Strzalkowski & Carballo, 1994, 1996; Strzalkowski, Carballo, & Marinescu, 1995; Strzalkowski, et al., 1997; Strzalkowski, Lin, & Perez-Carballo, 1998). The central idea is to augment the term frequency vectors with a frequency analysis of these richer features; text similarity measures between queries and documents are enhanced with similarity measures relying on these more sophisticated features.

Intuition might suggest that the application of richer linguistic features should be a clear win for text retrieval problems — what could be the drawback in using this additional information? Unhappily, any benefits seem to be often offset by problems (Lewis & Spark Jones, 1996). Indeed, the application of natural language features to text retrieval has generally floundered on one of two shoals: noise in the natural language processing system (e.g., the parser is not 100% perfect), or the added dimensions of these new features to an already high-dimensional problem further complicate the statistics. This is in addition to the simple fact that, as Lewis

and Spark Jones put it, the bag of words approach has already picked “the easy fruit off the tree” (1996, p. 11).

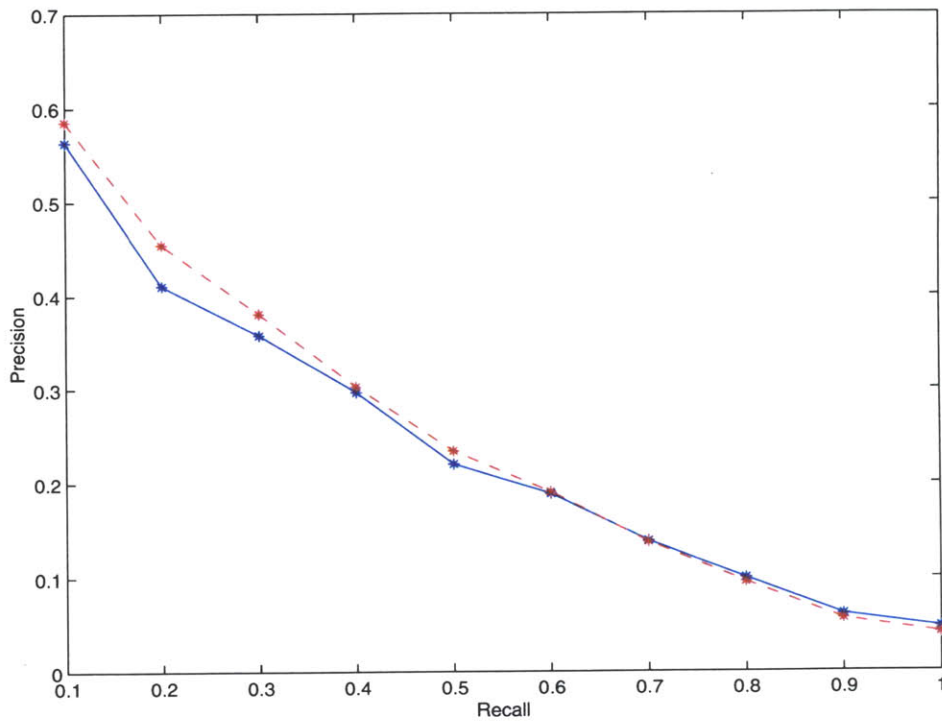


FIGURE 61. 10-point precision/recall experiment, averaged over 50 queries, with a standard corpus. Solid line represents retrieval using only term frequency information, dashed line includes syntactic phrase information. From Lewis and Croft (1990).

The above problems notwithstanding, over the years some limited success has been realized. Early experiments in applying parsing and phrase information to retrieval enjoyed improvements ranging from around 1% to 9% (Fagan, 1987; Smeaton & van Rijsbergen, 1988). Typical amongst these are the results depicted in Figure 61, due to Lewis and Croft (1990). Here, the solid line shows the 10-point precision/recall results (they do not compute precision for a recall of 0.0) using term frequencies only and averaged for 50 queries against a standard corpus. The dashed line shows the same retrieval tasks but with the engine augmented by a syntactic phrase system. The average improvement in precision due to the addition of phrase infor-

mation is 3.8%, which would not be considered noticeable under the Spark Jones categorization.

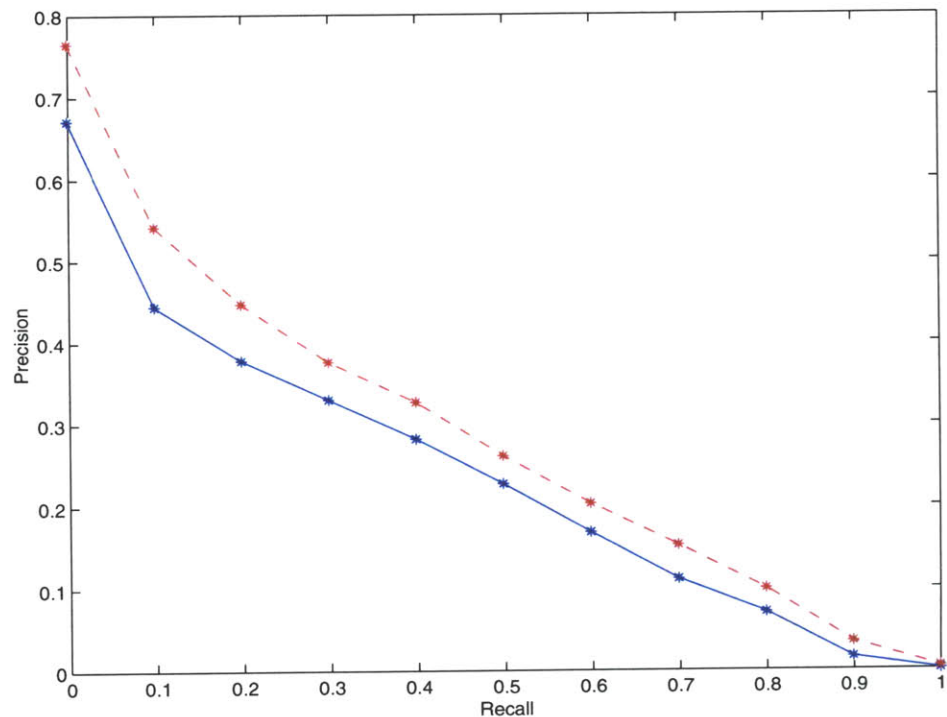


FIGURE 62. 11-point precision/recall results from TREC-3 ad hoc experiment. Solid line represents retrieval using only term frequency information, dashed line include syntactic phrase information and name recognition. From Strzalkowski, Carballo and Marinescu (1995).

More recent attempts to employ natural language techniques to aid retrieval have seen better improvements. Current research has been led, in particular, by Tomek Strzalkowski, along with a large set of collaborators (e.g., Strzalkowski & Carballo, 1994, 1996; Strzalkowski, Carballo, & Marinescu, 1995; Strzalkowski, et al., 1997; Strzalkowski, Lin, & Perez-Carballo, 1998). They have developed systems that demonstrate material improvement in retrieval by employing phrase extraction, name recognition, and other natural language technologies. In Figure 62, I show the 11-point precision/recall graphs for their system on the ad hoc queries from TREC-3. The solid line represents the average over 50 queries, for runs employing only

term frequency information. The dashed line adds phrase and proper name features. The increase in precision for this experiment, averaged over recall and queries, is 20%.

7.4 Results with Active Noun Phrase Replicators

I employed the CAMEL system as a retrieval engine, using standard term frequency analysis and the cosine similarity measure, for twelve queries against the Globe chronicon. The 11-point precision/recall graph was computed for each query.

I re-ran each of the twelve queries, this time augmenting the term frequency vectors with vectors representing the relative presence or absence of the active noun phrase replicators. This was accomplished by distilling from the Globe chronicon and the query text all noun phrases, using the methods described in Section 4.4.2. Each document, and the queries, were then scored, based on the relative presence or absence of those noun phrases deemed active (autocatalytic), by the methods of Section 3.10. Next, the distance, using the cosine metric, was measured between the active noun phrase vector for the query and the vector for each document. If a match between active noun phrase replicators was detected between a query and a document, then the distance between the query and the document, as measured by the traditional term frequency approach, was reduced by a weighted measure.

For example, “White House official” was determined to be an active noun phrase replicator within the Globe chronicon (see Table 20). If this lexico-syntactic replicator was detected in both a query and a document, the distance between the query and the document (computed as the cosine between their term frequency vectors) was reduced proportionally to the distance between the noun phrase vectors.

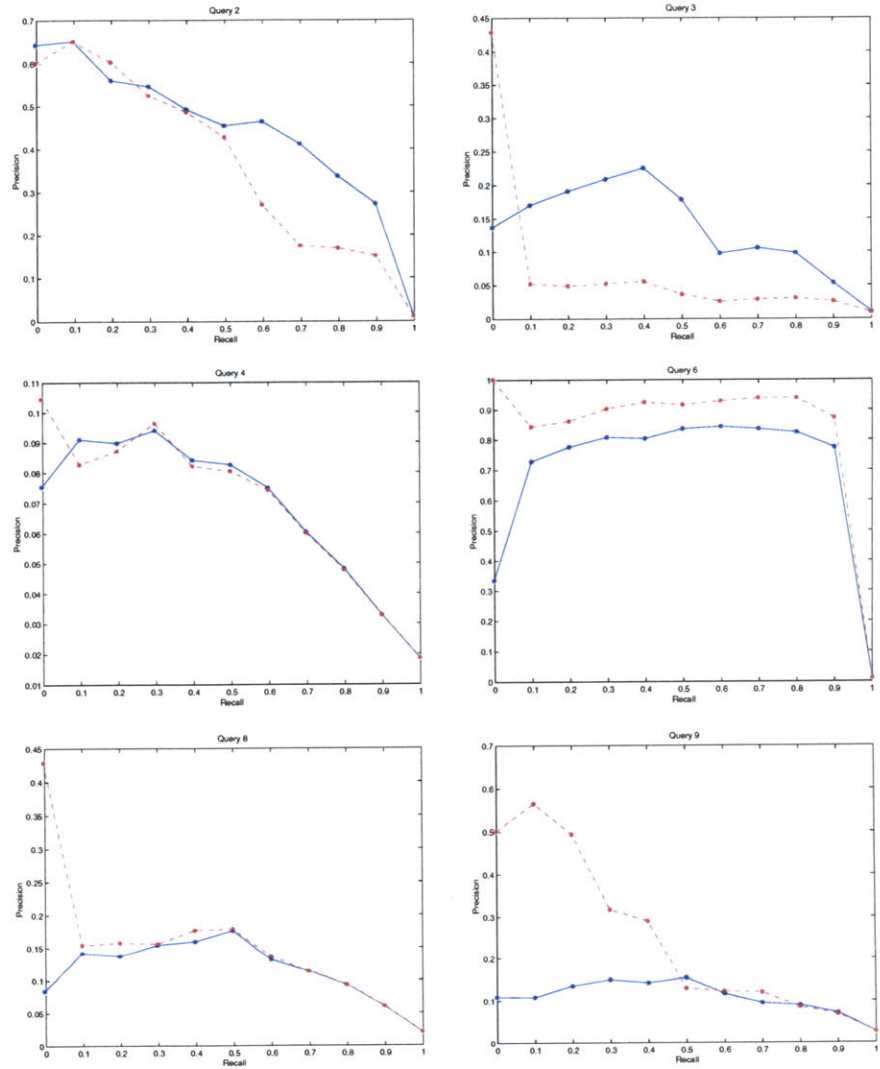


FIGURE 63. 11-point precision/recall graphs for queries that contained active noun phrase replicators. Results with terms only (solid) are compared against results with terms and noun phrases (dashed).

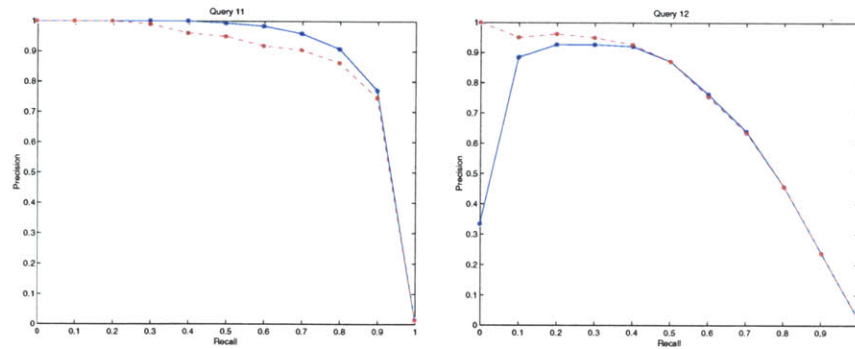


FIGURE 63 (Continued) 11-point precision/recall graphs for queries that contained active noun phrase replicators. Results with terms only (solid) are compared against results with terms and noun phrases (dashed).

In Figure 63, I show the 11-point precision/recall graphs for eight queries. The solid lines show the retrieval results using only term frequency information. The dashed lines show the retrieval results when the active noun phrase replicators were also employed.

Of the twelve queries, four of them did not contain any active noun phrase replicators. That is to say, the query text itself was free of any such replicators. In the absence of these features, there was no way to improve the retrieval, using this tech-

nique. In Figure 64, I show the single 11-point precision/recall graphs for these four queries, using only term frequency information.

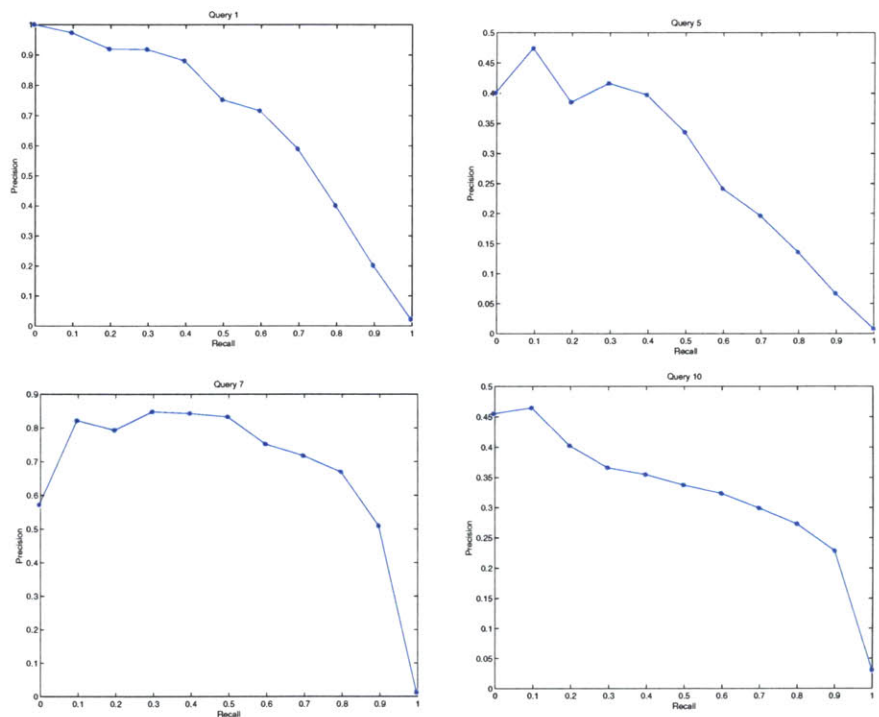


FIGURE 64. 11 point precision/recall graphs for queries that did not have any active noun phrase replicators.

7.5 Summary

The increase in precision due to natural language augmentation, averaged over the eleven points and twelve queries, was 8.2%. This is noticeable, but not material. However, the average increase for small recall (0.0 and 0.1) was 89%. It has been noted that measures at 0.0 recall can be misleading because they can be effected by a small number of irrelevant documents (R.K. Belew, personal communication, December 1999). That notwithstanding, I believe this result is worth our attention because improvements in precision for small recall are the most sought after. It is those first documents returned that a user is most likely to concentrate on. Such a material increase in precision for small recall suggests that this method could actu-

ally be employed to good effect in a real retrieval system. A practical strategy would employ the additional noun phrase features only for the first 10's of documents and then switch to using term frequency information. Figure 65 plots the 11-point precision/recall graph averaged over all queries; the improvement for small recall is quite noticeable.

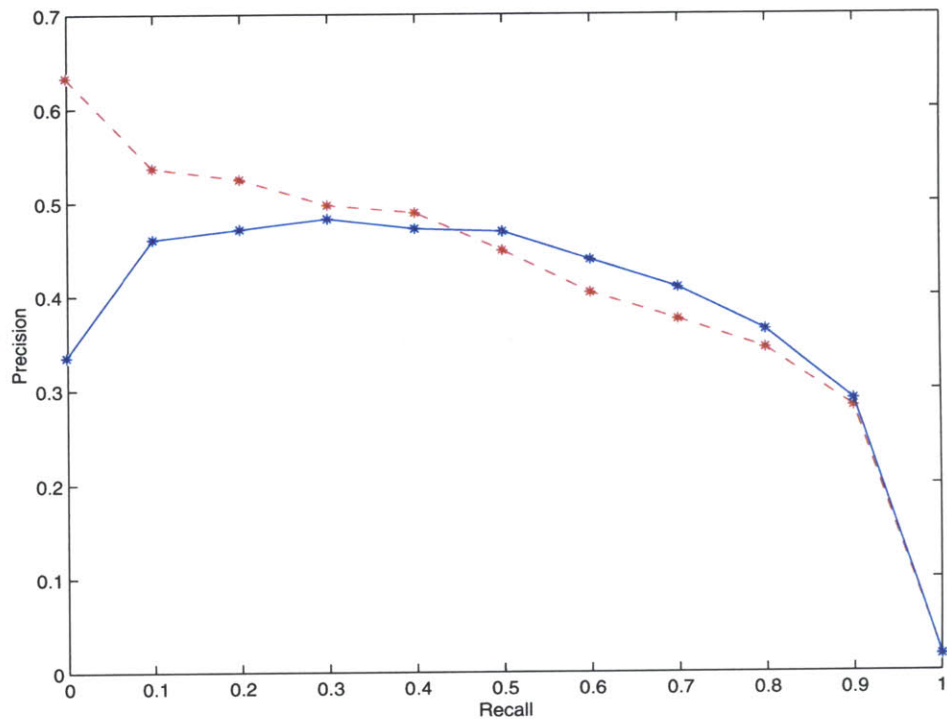


FIGURE 65. 11-point precision/recall results from *Globe chronicon* averaged over twelve queries. Solid line represents retrieval using only term frequency information, dashed line includes active noun phrase replicators. Noticeable improvement in precision is enjoyed for small recall.

Another obvious conclusion is that the current approach, employing active noun phrase replicators to aid retrieval, can only succeed when the query contains these replicators. Of the twelve queries, all of which were developed by people unaware of the nature of the retrieval experiment, four contained no such replicators. Because the system can detect the absence of these replicators, this is not such a

severe problem. In these cases, it simply does not add anything to the original term frequency approach.

Active noun phrase replicators can noticeably increase retrieval precision and materially increase it for small recall. This is the first NLP approach to retrieval that has made use of temporal data (which is encoded in the timeseries correlations). This approach competes reasonably well against other natural language informed retrieval systems that have been the result of a long research history. And, indeed, for small recall, this approach beats the best systems which have emerged from multi-year focused research programs involving very large collaborations between GE, Lockheed Martin, Rutgers, and NYU (Strzalkowski, et al., 1997).

This dissertation crosses a number of intellectual traditions and research communities and, as such, the related work spans a broad range. The primary research communities and related work to this dissertation have already been discussed in the earlier chapters, for instance, evolutionary theory (e.g., Williams, 1966), corpuslinguistics (e.g., Sinclair, 1991), information retrieval (e.g., Salton & McGill, 1983), and natural language processing (e.g., Karlsson, Voutilainen, Heikkilä & Anttila, 1995).

In this chapter I will examine some of the wider research communities whose work has impacted this dissertation. Those wider disciplines include evolutionary models of culture, in general, and language, in particular; computer based simulations of culture and language evolution; memetics; language change theories; and studies of internet discussions. I will start with the most general investigations into evolution and culture, narrow to evolution and language, and narrow again to investigations of language dynamics and retrieval within collections of text. I end with a quick review of some ongoing work here at the MIT Media Laboratory.

8.1 Evolution and Social Behavior

Quite a number of researchers, from psychologists to evolutionary theorists to behavioral ecologists, have explored cultural and social behavior within an evolutionary context. Within human society it has been argued that “language is the best

approximation of pure culture” (Gerard, Kluckhohn & Rapoport, 1956). What is true for language is true, *a fortiori*, for other areas of culture. Therefore, my results in language microevolution should speak to cultural evolution at large and, vice versa.

8.1.1 Evolutionary culture theories

A large community of researchers have developed theories of transmission and evolution of cultural traits within a neo-Darwinian framework. Some substantial theories of cultural transmission and evolution include Lumsden & Wilson (1981), Cavalli-Sforza & Feldman (1981), Boyd & Richerson (1985), Barkow (1989), and Durham (1991). Durham (1990) does a nice job of reviewing a wide range of evolutionary culture theories.

Three of these models are particularly worth our attention. Cavalli-Sforza & Feldman enumerate a variety of forces that come to play in the formation of cultural traits: “(1) *mutation*, which is both purposive (innovation) and random (copy error); (2) *transmission*, which is not as inert as in biology; (3) *cultural drift* (sampling fluctuations); (4) *cultural selection* (decisions by individuals); and (5) *natural selection* (the consequences at the level of Darwinian fitness)” (Cavalli-Sforza & Feldman, 1981, p. 351, emphasis in original). Their work, however, concentrates most on (2): the role of transmission forces (e.g., model-to-learner ratios). Boyd and Richerson (1985) also concentrate on transmission structures in their “dual inheritance theory” of genes and culture. They argue that culture undergoes an evolutionary process via an inheritance system that is structurally different from the genetic system and can at times work in conflict with it (1985, p. 2). Durham (1991, chap. 7) also offers a coevolutionary theory under which culture and genes can be in “opposition” to each other.

All of these works offer a critique (at least implicitly) of simplistic sociobiological models wherein culture is kept on a “leash” by genes (e.g., Wilson, 1978, p. 167).

Ongoing research has continued to explore the relationship between cultural and genetic systems and modes of interaction between them. Kevin Laland and co-authors (Laland, Kumm & Feldman, 1995) propose a test case in gene-culture coevolutionary theory. Their analytic model considers cultural factors that influence human sex ratios, such as “sex-selective abortion, sex-biased infanticide, [and] sex-prejudicial abandonment” (p. 135). They model how these cultural factors can impact potential genetic biases for producing one or the other sex.

Mark Feldman and Laland (1996) also produced an important review paper on gene-culture coevolutionary interactions. They consider such diverse phenomena as lactose absorption, female mortality, and the spread of agriculture. And Odling-Smee has joined them both on an upcoming target article in *Behavioral and Brain Sciences* (Laland, Odling-Smee & Feldman, in press). Here, they review their theory of niche construction in which an organism's behavior, including socio-cultural aspects, impacts its offspring. Cultural behavior acts as a form of extra-genetic inheritance: "Parents in a vast numbers of species, across broad taxa, act in ways that influence the developmental environments of their offspring, for example, by providing them with benign nest environments or with food" (Laland, Odling-Smee & Feldman, in press, section 1.1).

In a recent paper, Lachlan and Slater (1999) have developed coevolutionary models of the development and maintenance of vocal learning in song birds. They argue that cultural evolutionary forces can drive the gene-culture relationship into a "trap" that sustains certain cultural processes but, perhaps, is not optimal when viewed solely from the gene's vantage. This is one of the few works to explore gene-culture coevolution in nonhuman animals.

8.1.2 Evolutionary psychology

A research program related to evolutionary culture theory, and at times employing a coevolutionary approach, has sought explanations for cognitive aspects of human behavior within a neo-Darwinian framework. Working under the name "evolutionary psychology," this programme has enjoyed the inspiration and leadership of, in particular, researchers Leda Cosmides and John Tooby.

Cosmides and Tooby set the stage for evolutionary psychology by asking not what the consequences of some behavioral property of humans is today but what its adaptive evolutionary history was under Pleistocene conditions (Cosmides & Tooby, 1989; Tooby & Cosmides, 1989). With this radical question they propose to undermine the "Standard Social Science Model" under which humans are born with a "general-purpose, content-free psychology" and "biology is intrinsically disconnected from the human social order" (Tooby & Cosmides, 1992, p. 49). In a wide range of experiments (e.g., Tooby & Cosmides, 1989; Cosmides & Tooby, 1992), they have developed a social exchange theory which examines how human psychologies process social contracts, such as, "if you give me *P* then I'll give you *Q*" (Cosmides & Tooby, 1992, p. 80). They offer evolutionarily informed explanations as to how conditions in the Pleistocene may have led to the development of specialized cognitive modules that handle these social contracts.

A number of recent books have sought to collect important papers within evolutionary psychology (Barkow, Cosmides & Tooby, 1992; Betzig, 1997), review, reveal, and expand on important elements to the theory (Pinker, 1997), and popularize the research programme (R. Wright, 1994).

8.1.3 Social learning theory

A wealth of fascinating work has emerged from the experimental psychology and behavioral ecology communities studying social learning and cultural transmission amongst, in particular, nonhuman animals. While this community has not always looked for evolutionary explanations under which communities engage in social activities, they are at the forefront in exploring mechanisms and processes for social transmission. A collection of significant research projects have been underway, some for twenty or more years. A nice and reasonably recent review of this work can be found in the collection of papers edited by Heyes and Galef (1996).

The major players and projects within the research community include: Jeff Galef (e.g., 1994) who has been studying food preference learning via social transmission in the Norway rat (*Rattus norvegicus*); Luc-Alain Giraldeau & Louis Lefebvre (e.g., 1987) have looked at the cultural transmission of foraging behavior in pigeons; Cecilia Heyes (e.g., Heyes & Dawson, 1990) has studied imitation (or lack thereof) in rats via a bi-directional experiment. Rats can receive a reward by moving a joystick to the left or right; Heyes and co-authors have found that observer rats have a tendency to move the joystick in the same direction as a trained model does; Andrew Whiten (e.g., Whiten & Cusance, 1996) has developed an “artificial fruit” which consists of a transparent box with a reward inside it and a complex latch which secures its lid. Whiten and colleagues have examined imitation in chimps by studying how they apply observational learning to open the artificial fruit and gain the reward. (Whiten has also recently led a fascinating comparative study of chimp cultures (Whiten, et al., 1999).)

Researchers studying social learning in nonhuman primates form a significant community in their own right (e.g., Huffman, 1996; along with Whiten & Cusance, 1996).

8.1.4 Memetics

Richard Dawkins (1976), in a lovely turn of phrase, proposed “meme” as an analog to gene. A meme is a replicator within an evolving socio-cultural environment. Dawkins intent was to illustrate the generality of the concept of replicator — that it

was not just relegated to organic, genetic evolution. To his surprise a community of researchers have emerged, calling themselves memeticists, who study evolution and transmission of cultural replicators.

While Dawkins is credited with the neologism “meme” the term does indeed have many antecedents. The German researcher, Richard Wolfgang Semon (1908/1921), published at the turn of the century *The Mneme*, a book which described a sort of memetics. Semon claimed to “discover analogies between the various organic phenomena of reproduction... and the other kind of reproduction which we call memory” (1908/1921, p. 9; see also Hull, 1999b).

Since Dawkins’ short chapter on memetics, a number of important thinkers have taken on the idea, including most notably Dennett (1995) and Hull (1982). A recent conference at Cambridge University assembled most of the major players within this research community (see Aunger, 1999). An outcome of this meeting was the observation, as Hull (1999b) put it, that “the clock is ticking”: The memetics research programme must demonstrate that it is a progressive endeavor, or it will wither away much as have other failed scientific projects.

Besides the works of Hull, Dennett, and Dawkins, a few other researchers have offered high-level treatments of memetics. These include Percival (1994), Brodie (1996), Lynch (1996), Tracy (1996), J.S. Wilkins (1998), and Blackmore (1999). The only work to explicitly consider memetics within written natural language, and thus the work most closely related to my own, is that of Elan Moritz (1990). However, this paper offers only a high-level theoretical treatment.

A few researchers have attempted to actually *do* memetics; that is, they have assembled empirical support of a cultural replicator theory. But, given the volume of work published within this general programme, empirical studies are by far the exception. Those few studies have included an examination of human cultural dynamics (Deb, 1996), human business and policy making dynamics (Speel, 1997; de Jong, 1999), and the analysis of social dynamics amongst song birds of Western Australia (M.C. Baker, 1996). In particular, the contribution of Debal Deb (1996) offers a nice example of what can be done with empirical population memetics. Deb studied the transmission, variation, and maintenance of techniques for net fishing between two Indian caste groups. He found that certain fishing communities maintained an overhead casting technique when fishing, even though a waist-level method was easier to learn and more efficient. Deb argues that the overhead technique was maintained as a cultural trait in order to differentiate between castes.

Related Work

A larger group of researchers have employed computer simulation and modeling to explore memetic dynamics. This includes the work of Bura (1994), Gabora (1995), and Hales (1998). These simulation environments often bring together such a variety of complex dynamics that it can be difficult to draw solid conclusions from their results. For instance, Gabora (1995) evolves neural networks — a system notorious for its baroque complexity. (Similar complaints can be lodged against some of the computer simulations of language evolution described in Section 8.2.4.)

Some more modest computer simulations have attempted to study simpler memetic dynamics. In particular, the relationship between individual learning, organic evolution, and cultural replicators was considered by Belew (1990) and myself (Best, 1998b, 1999b, in press).

Bruce Edmonds (1998) has compared and contrasted the various simulation models within memetics concentrating on my own work (Best, 1997) and the more abstract models of William Calvin (1997).

Plenty of controversies and, frankly, invective has arisen within the nascent memetics community. Some of this is considered in Rose (1998), Gatherer(1998), and Best (1998c), along with published responses. Rose cites the main controversies as “ambiguity in the definition of a meme and confusion regarding the distinction between replicator and phenotype, the problem of inheritance of acquired characteristics, the relationship between memetics and sociobiology, and the selection or mutation of memes being carried out by conscious foresight” (Rose, 1998, paragraph 1). I’ve made more general criticisms, noting an immaturity in formal modeling and lack of empirical support to most work. I also suspect many controversies come from a tendency to either over-read Dawkins (e.g., in terms of the importance of true imitation (Blackmore, 1998, 1999) or neural storage of information (Lynch, 1998)) or under-read Dawkins (e.g., in terms of identifying an appropriate unit of selection (Gatherer, 1998)).

The work reported in this dissertation offers, in my estimation, the first complete model of population memetics. By that, I mean it provides an operational definition of a cultural replicator within a formal model along with supporting empirical results.

8.1.5 Controversies in cultural evolution

Contemporary theories of cultural evolution have received considerable criticism. This is in addition to earlier well deserved attacks on simplistic models of sociobiology and Social Darwinism (Caplan, 1978).

Stephen J. Gould, the essay-writer-in-chief for the North American evolutionary community, has led the attack against contemporary evolutionary culture theories. Gould (e.g., 1991, 1997a, 1997b) attacks evolutionary theories of culture and evolutionary psychology for an over reliance on natural selection and single-mindedly searching for (and finding) adaptations. Much of the time Gould is attacking rhetorical strawmen — caricatures of modern evolutionary culture theories. Further, he confuses the ultra-Darwinist with Universal Darwinism. In other words, he fails to understand that the *pervasiveness* of the evolutionary algorithm can nonetheless admit to a *plurality* of explanation and cause within specific instances.

The geneticist Steve Jones has labelled Gould “a snail geneticist gone to the bad” (Brockman, 1995, p. 70). Jones joins a group of researchers who have sought to attack and undermine Gould and his criticisms of cultural evolution (Pinker, 1994; Dennett, 1995).

8.2 Language

From a consideration of culture at large, I shall now narrow focus to a singular aspect (or engine) of culture, namely, human natural language, and review evolutionary explanations of it. This section ends with a description of some of the computational and formal models of language evolution.

James Hurford has noted that “the phrase ‘evolution of language’ carries an unfortunate ambiguity. It can be understood as describing the glossogenetic processes, studied by historical linguistics, whereby actual languages... evolve into daughter languages.... Alternatively, it can be taken to describe the phylogenetic processes by which the capacities of our remote ancestors... evolved into the language faculty innate in modern *Homo sapiens*” (1991, p. 273). Section 8.2.1 reviews work on the phylogenetic evolution of language, while Section 8.2.2 considers glossogenetic evolution.

It is worth noting that Hurford’s description of the existing two types of language evolution research ignores the possibility of studying language microevolution.

Indeed, this dissertation may well suggest a third track within the “evolution of language” programme — microevolution.

8.2.1 Phylogenetic language evolution

Liberation, according to James Hurford (1990; see also Aitchison, 1998), came in the form of a target article in *Behavioral and Brain Sciences* by Pinker and Bloom (1990). In their article Pinker and Bloom argued, contra Gould and Chomsky, that the human language faculty can be explained as the outcome of a neo-Darwinian evolutionary process. This article, along with a book-length synthesis of current ideas in evolution and language (Pinker, 1994), has helped to rehabilitate the phylogenetic language research programme.

Lieberman (e.g., 1992, 1998) has developed a large body of research on the evolution of language faculties. He has researched language evolution starting from the basic physiology of speech and vocal control and moving all the way to the selective advantages of syntax. Other treatments include W. Nobel and Davidson (1996) who have, in particular, studied the emergence of symbolic languages. Robin Dunbar (1996) has advanced a substantial theory of the evolution of the language faculty. He argues that language is the outcome of early hominids reliance on gossip to maintain social cohesion. This was in substitution for grooming, which non-human primates rely on to establish and sustain the same social bonds. These works are complemented by a range of other studies on phylogenetic theories of language evolution (e.g., Hurford, 1992; Gyöi, 1995; Hildebrand-Nilshon, 1995).

Some researchers have studied the evolution of language faculties for non-humans. In particular, non-human primates have been studied, including Locke and Hauser’s (1999) exploration of vocal signaling in vervet monkeys. Recent work has reported evolutionary explanations for the origin of the honeybees’ waggle dance (Dornhaus & Chittka, 1999).

Some recent collections of papers focusing on phylogenetic language evolution include MacWhinney (1999) and Hurford, Studdert-Kennedy and Knight (1998).

8.2.2 Glossogenetic language evolution

The study of glossogenetic evolution is a mammoth and long-lived research endeavor that encapsulates much of historical or comparative linguistics. Since it is beyond the scope of this dissertation, I am not able to do justice to this literature here. Instead, I will simply touch on the relationship of historical glossogenetics to

Darwin and his contemporaries and the relationship of language change to glossogenetic evolution and the active replicator model advanced in this dissertation.

Much of 19th Century linguistics centered on glossogenetic theories which were informed by 19th Century evolutionary theories. The concept of “evolution,” pre-dating Darwin, referred to some general transformation, progress, or decay, and glossogenetic research concentrated on revealing the *scala naturae* of the Earth’s languages (Stevick, 1963; Nerlich, 1990; McMahon, 1994). Darwin was aware and made use of these early evolutionary theories of language. Contemporary with the publication of his *On the Origin of Species* (Darwin, 1859/1964), other researchers looked for ways to inform their studies of language change with a specifically Darwinian view of natural evolution. For instance, Charles Lyell (1863) devoted a chapter of his book on uniformitarian principles for geological change to the consideration of historical language evolution, describing it as the accumulation of slight variation resolved through selection due to preference. August Schleicher wrote in 1868, soon after the publication of *Origin of Species*:

Languages are organisms of nature; they have never been directed by the will of man; they rose, and developed themselves according to definite law; they grew old, and died out. They, too, are subject to that series of phenomena which we embrace under the name of ‘life’. The science of language is consequently a natural science...

(as cited in Keller, 1994, p. 48)

And Arsène Darmesteter (1886) argued that words struggle for survival within individual psychologies leading to the historical evolution of language. See Nerlich (1990) for a review of this history.

Moving to the 20th century, any theory of glossogenetic evolution has required an understanding of the agents of variation and change. Within the linguistics community, little has been made of language change in the 20th century. What does exist of recent work has primarily viewed language change as a subject of historical or comparative linguistics (e.g., Arlotto, 1972; Jeffers & Lehiste, 1979; Anttila, 1989; Polomé, 1990; McMahon, 1994; Lass, 1997). In other words, language change theories have rarely been within an evolutionary context and never a microevolutionary context. (In Section 5.6, I cite much of the related work specifically on lexical semantic change.)

However, a few researchers have concentrated on language change without looking at it strictly in terms of history-scaled language development or comparative linguistics. For instance, Keller (1994) studies language change from an economic perspective. Some resurgence of interest in language change has been prompted by the growth of the sociolinguistic community (Labov, 1972). This research programme is especially interested in the rise of non-standard varieties (in particular, sound innovations) within identified social communities.

In a fascinating and influential paper Weinreich, Labov, and Herzog (1968; see also Bauer, 1994) proposed a collection of Hilbert questions for a theory of language change: (1) The constraint problem — what are possible types of change and what are the possible conditions for change. (2) The transmission problem — how does a linguistic system move from one state to another, is it gradual or abrupt, regular or irregular. (3) The embedding problem — what other changes are associated with the given changes in a manner that cannot be attributed to chance. (4) The evaluation problem — what is the effect of change on the linguistic system or upon the efficiency of the communication system. Many, if not all, of these questions are available for exploration under my active replicator model, and some of the results of Chapters 4 and 5 may shed light on these very questions.

8.2.3 Critiques of the phylogenetic/glossogenetic split

A number of researchers have criticized the phylogenetic/glossogenetic dichotomy. Many papers from the collection edited by Puppel (1995) fault this distinction arguing, in particular, that a language faculty and languages over time are co-adapted to each other (Bichakjian, 1995; Messer, 1995; Smillie, 1995). Terrence Deacon (1997) produced a book-length argument on the co-evolutionary relationship between languages and the language faculty. He reasons that languages over time have evolved to be better suited to their environment, namely, the information-processing landscape described by the human language faculty.

A couple of collections have reported on results within both the glossogenetic and the phylogenetic programmes of language evolution (de Grolier, 1981; Hawkins & Gell-Mann, 1992).

8.2.4 Computer and formal models

A number of labs have focused recently on computer simulations and formal modeling of language evolution. This work has helped to constrain the algorithmic complexity of various language phenomena — how simple can a system of interacting

agents be and still allow for the emergence of a language system. Most of these systems place in simulation a collection of agents who interact within some shared environment. Various language skills are developed or honed via the genetic algorithm, sometimes in conjunction with a neural network.

A particularly popular area of study by simulation has been the evolution of shared meaning and lexicon formation (Werner & Dyer, 1992; MacLennan, 1992; Ackley & Littman, 1994; Steels, 1996; Arita & Taylor, 1996; Noble & Cliff, 1996; Saunders & Pollack, 1996). While I find these studies of considerable interest, many of these simulation environments are so complex as to obscure their end result. Other computer simulations have studied innateness and critical periods (Hurford, 1991; Batali, 1994), the evolution of phonology and vowel systems (de Boer, 1997), emergent grammar and structure (Hashimoto, 1994), and computational models of composability as a complex emergent property of social transmission (Kirby, in press).

A large body of work has emerged from the lab of Luc Steels. His group has been studying language as an emergent complex system due to “evolution, co-evolution, self-organization and level formation.” They are “exploring this hypothesis in a series of experiments on robotic and software agents that span all aspects of language: grounded meaning creation, lexicon formation, syntax, and emergent phonology” (Steels, 1996, p. 562). Some of Steels’ recent journal publications are indicated in the references (1998a, 1998b).

Much of this computational simulation work in the evolution of language has been reviewed by Bill Turkel (1997).

A couple of researches have developed quite simple computer models of lexical semantic change (Clarke & Nerlich, 1991; Mair, 1997). Neither of these simulations, however, explicitly considers the evolutionary consequences of language change. This work underlines the considerable potential in computational analysis of language change, but neither makes significant progress towards that goal.

Some researchers have explored formal mathematical modeling of the evolution of communication. Niyogi and Berwick (1996) have examined the principal-and-parameters theory and shown how many positive examples are required (only 100’s) to set a parameter correctly with high probability. Nowak and Krakauer (1999) have studied game theoretic formal models of the evolution of protolanguages within a nonlinguistic society, in particular, the evolution of lexicon formation. And Worden (1998) has proposed a unification-based grammar which

represents word meanings and allows structure (syntax) to develop as an emergent property.

8.3 Text

I will now move from language in general, and the evolution of language in particular, to general methods of studying texts, usually outside of any evolutionary model. Most of the related work in these fields has already been reviewed, so I will only gloss some of the more ancillary works here.

8.3.1 Corpuslinguistics and text analysis

The last twenty years has seen an incredible increase in results from the corpuslinguistics community. This is due primarily to the rise in available computational power and the extraordinary increase in online corpora. The corpuslinguistic work most related to this dissertation examines lexical phenomena.

Researchers have explored lexical variation through large collections of texts (Oostdijk, 1988; Kjellmer, 1994). These studies rely on fairly simple stylostatistical measures such as word frequency distributions and word length distributions; I used related approaches when comparing the Clinton and Globe chronica in Section 3.13. Some richer analysis methods have been proposed by Pustejovsky, Bergler and Anick (1993). They attempt to study lexical semantic variation in a corpus by proposing a set of semantic descriptions for each word.

Some researchers have employed similar corpuslinguistic techniques for specific problems such as lexical disambiguation (Gale, Church & Yarowsky, 1993; Asher & Lascarides, 1995) or topic detection and discovery of content bearing words (Bookstein, Klein & Raita, 1995; Allan, Carbonell, Doddington, Yamron & Yang, 1998). I have found that autocatalysis is a good indicator of words or phrases that represent topics and bear heavily on content (Section 4.6). These related approaches rely on word frequency and clustering and none make use of temporal data. Important early corpuslinguistic studies of keyword identification and thesaurus construction were performed by Karen Sparck Jones (1971). A more recent substantial investigation into thesaurus construction can be seen in Grefenstette (1994).

In Section 3.1, I mentioned that manuscript traditions (e.g., Chaucer's *The Canterbury Tales*) form a type of chronicon, as they describe a collection of texts over time. Some interesting corpuslinguistic investigations have applied cladistic analysis, as would be used to describe historical relationships within systematic biology,

to piece together the historical links between documents within a manuscript tradition. Robinson and O'Hara (1996) have reported on their use of such techniques in developing a history of Old Norse sagas. This offers complementary examples of work where corpuslinguistic techniques and biological theories inform an analysis of texts over time.

Collocations are cohesive clusters of words that co-occur together more than would be expected from chance (Smadja, 1993). The analysis of collocations is an important related field, as it attempts to identify appropriate sized replicators in a way that is not dissimilar to an analysis of the size of units of selection (see Section 5.8).

A number of useful general collections and books have been published recently on the topic of corpuslinguistics. Aarts and Meijs (1990) has a nice collection of general papers on the topic, while Zernik (1991) concentrates on phenomena at the lexical level. Oakes (1998) has written a very accessible general introduction to computational corpuslinguistic techniques, as has Barnbrook (1996).

8.3.2 Text retrieval

The CAMEL software system is responsible for the principal results that structured this dissertation. In organization and function, CAMEL resembles a text retrieval and natural language processing system. In this section, I will touch on some of the state of the art techniques for text retrieval that relate to or contrast with the CAMEL system.

The prime goal of the CAMEL software system is to distill linguistic replicators. A number of text retrieval techniques have employed similar analyses to aid retrieval. This has included clustering terms and phrases that reoccur frequently across a text collection and using these clusters as features for retrieval (Lewis & Croft, 1990). A simpler method is to group together collocated words into n -length windows; this is similar to the approach I took when studying syntactic n -grams (Section 4.5). Cavnar (1975) used n -length strings of words as features for retrieval. The benefits to retrieval from these two techniques have been mixed.

The Microevolutionary Language Theory focuses on complex design at the simplest levels, such as the accumulation of lexical semantic innovation. A number of researchers have applied lexical semantic knowledge to the retrieval task. Schütze and Pedersen (1995) used similarity, determined through lexical co-occurrence, to disambiguate word senses. Each word sense would then serve as a unique feature for retrieval, resulting in 7% to 14% improvements in performance (in the sense

described in Chapter 7). Similar approaches have been explored through the application of WordNet, a well known lexical database (Miller, 1990). For instance, Voorhees (1998) attempted to use WordNet to aid retrieval. However, the inability to correctly disambiguate word senses limited the effectiveness of this approach.

A number of researchers have applied evolutionary algorithms, especially the genetic algorithm (GA), and other learning techniques directly to text retrieval. Pattie Maes and co-authors have developed computer systems to evolve text analysis and retrieval agents (Sheth & Maes, 1993; Moukas & Maes, 1998). These agents act as personal information filters monitoring information published on the web, NetNews, or email and tagging relevant documents. Menczer and Belew (2000) developed retrieval agents that search links on the web and learn, through evolutionary adaptation, to identify pages relevant to a user. Yang and Korfhage (1993) use the GA to optimize the queries sent to a text retrieval engine.

Hsinchun Chen (1995) wrote a survey of AI approaches, such as GAs, neural nets, and symbolic learning, in information retrieval applications. He identifies a number of pros and cons for each approach. A more recent contribution from Richard K. Belew (2000) reviews and synthesizes many AI approaches to information retrieval.

8.4 Media Laboratory Work

A number of other Media Lab researchers have explored discourse, such as NetNews, over the Internet. Ongoing work at the Lab includes a collection of projects which analyze and visualize posts to NetNews (Karahalios, 1998; Xiong, 1998). An interesting set of studies has focused on detection of point of view (Sack, 1994) and of “flames” or abusive messages (Spertus, 1997) over Internet discourse (this work came primarily from the MIT AI Lab). Warren Sack (1999, in press a, in press b) has used parsing and text analysis techniques to track discourse elements and relate them to their authors. His Very Large Scale Conversations (VLSC) are fundamentally the same as my chronica; they both describe collections of texts over time with multiple authors and readers.

A number of Lab researchers have used and improved on text retrieval engines. Brad Rhodes (Rhodes & Starner, 1996) has been exploring text retrieval methods to deliver just-in-time information. In this system, software agents continuously analyze a user’s work context (e.g., the email currently being read) and propose documents that may be relevant. Daniel Gruhl (2000) has developed text systems that employ a range of feature extraction techniques, including natural language processing approaches, to aid retrieval.

The most closely related work to this dissertation is my own. In earlier work, I employed the strictly statistical singular value technique first proposed by Deerwester, et al. (1990). With only statistics, I was able to make surprisingly good progress on a number of theoretical issues in the microevolution of text including the identification of units of selection (Pocklington & Best, 1997; reviewed in Section 5.5), average population fitness (Best, 1998a; reviewed in Section 3.10), adaptive value (Best, 1999b; reviewed in Section 4.3), and competition (Best, 1997; which forms most of Chapter 6).

8.5 Summary

This work crosses quite a number of disciplines — from text analysis to corpuslinguistics, evolution of language to evolution of culture. Indeed, some of its value may lie with its particular integration of contemporary replicator-based theories of evolution with contemporary theories and practice of corpuslinguistics and text retrieval.

In this chapter I have sought to review some of the main works across these related disciplines. In the next chapter I end by arguing why this set of theories and practice belong together, and how my work relates to and contrasts with them all.

Related Work

Bad luck with biological models has left historical linguistics with such a heritage of confusion and specious explanations as to condition linguists to reject or ignore all putative parallels between languages and living organisms (Stevick, 1963, p. 159).

Stevick expresses the attitude, and the outcome, of many modern attempts to bring together biological and evolutionary theories with linguistics. Indeed, much of modern linguistics, as dominated by Chomsky, has set biology at quite a distance from human natural language.

I believe, however, that the 1990's has seen a significant turnaround of luck for this programme of conceptual integration. Pinker and Bloom's (1990) *BBS* article may serve as a marking point for this change in fate; in the last ten years, we have seen a wealth of successful attempts to integrate evolution with explanations of human natural language. And now, at the dawn of a new millennium, we have new found interest and enthusiasm in these avenues of exploration as evidenced by new and ambitious conferences, journals, and books.

By the end of the next decade, we will see evolutionary theory positioned as a cornerstone to the explanation of how humanity developed its general language competency, and furthermore, how each language, as used, has developed its unique range of particularities.

I will conclude this dissertation by asking: What is it good for? How is it different? and, Where can we go from here?

9.1 What is it good for?

I have proposed an active language replicator model and used it, along with a toolkit of text analysis software, to build support for a Microevolutionary Language Theory. But what is all this good for?

The Microevolutionary Language Theory strives to:

- identify the where and wherefore of complex design elements within human natural language;
- answer the *Cui Bono?* question within natural language — active language replicators are the central beneficiary of the evolutionary process;
- provide insight, given a better understanding of the macroevolutionary consequences of microevolution, to the large-scale patterns of historical linguistics;
- support and elucidate Campbell's Rule and Universal Darwinism.

Underlying the Microevolutionary Language Theory is an active replicator model and a set of text analysis tools. This model and software system strives to:

- track language (and by way of language, culture at large) over time and identify what's hot and what's not;
- support comparative studies of language across texts, time, media, and communities;
- provide a toolkit to support empirical observations of population memetics and focus on the appropriate units and targets of selection;
- improve on practical problems of text analysis such as retrieval.

9.2 How is it different?

In a number of ways this dissertation describes work that is novel and an improvement, I hope, on past research endeavors.

There has been considerable related work on the evolution of culture and, in particular, the evolution of language; in Chapter 8, I reviewed much of this work. My work is distinct in its strong replicator or microevolutionary focus and in its attention to identifying appropriate units of selection. A vague conception of the units of selection has been the source of much confusion in the evolution of culture and language. Strong examples of this sort of confusion appear within the memetics com-

munity where, for instance, “religion” or “agriculture” are often treated as targets of selection in their own right. A project to model the transmission, variation, and selective retention of “religion” seems doomed to failure (in contrast, say, to modeling some particular tenet or liturgical trait). One important way that this dissertation relates, and contrasts, with existing studies on the evolution of culture and language is that it attends directly to the problem of identifying appropriate units of selection and proposes an operational, engineering solution to this problem (namely, a big software system that distills simple active replicators).

Most all studies of cultural or language evolution have been formal, computational, or philosophical. Few have backed up their models with significant empirical studies. Online text collections and text analysis systems, as employed by my study, are on the increase and provide an ideal and ample set of empirical data to work from. In this work, I was able to support my theory with hundreds of empirical observations from text collections.

Studies into the evolution of language have concentrated on either the development of human language faculties or on the historical evolution of languages over long periods of time. Little or no work has been devoted to the study of the microevolution of language at the simplest levels. Those researchers that have studied change in language at the simplest levels (e.g., lexical semantic change) have not made use of an evolutionary perspective and, thus, have missed potential, explanatory avenues. Here, I have examined simple language change and innovation with a microevolutionary model and, further, have proposed ways in which microevolution might relate to these other forms of language evolution at differing scales.

This dissertation relies on a large number of techniques from the natural language processing and text retrieval communities. It contrasts with related work most substantially in its reliance, thanks to an evolutionary focus, on temporal activity. For instance, text retrieval is improved by employing natural language techniques along with temporal data (which is how active replicators are distilled). This particular temporal approach has not been previously used in text retrieval.

In contrast to the text retrieval community, the corpuslinguistic community has made use of time, especially with many of its stylostatistical measures. My work is distinct, particularly, in its focus on internet discourse and collections. These communities of discourse on the net are an important environment to study.

9.3 Where can we go from here?

A range of future work would go to improving and extending this research. This includes studies to further support the Microevolutionary Language Theory in general, and improvements to the text analysis system in particular. Further, this work suggests many follow-up experiments extending it in new ways.

9.3.1 New work in support of the Microevolutionary Language Theory

One obvious (and important) way to add support for the Microevolutionary Language Theory is simply to accumulate more examples of active replication across more genres and collections and over longer time periods. In particular, it would be useful to demonstrate complex design, such as was attempted with the “Nazi” replicator in Section 4.3.3 and Section 5.6, with a wide range of examples. A particularly powerful example would capture some language innovation at its birth and track its utilization, demonstrating microevolutionary pressures, until it was pervasive in the community of use. (For instance, could we track the arrival and acceptance of a phrase such as “surf the web”?)

One of the most time consuming and labor intensive elements in developing more examples of this sort is the difficulty in acquiring and assembling appropriate chronica. A worthy future project would assemble a wide collection of texts appropriate to these sorts of studies and share them with all interested researchers.

Besides further examples of complex design, the Microevolutionary Language Theory would be enhanced by a better understanding of the mechanisms of transmission and copying and the causal links between a linguistic trait and its subsequent reappearance. Understanding these dynamics requires, to a large extent, tracing of the cognitive or psychological processes of the individual authors (in other words, the role of the author is the key to transmission mechanisms and causal links). An open question is to what extent experiments that rely solely on an analysis of the text chronicon might be able to answer questions of transmission and causation. A cleverly designed experiment could perhaps distill these processes, or surrogates of them, from the text only. Currently, however, no convincing experiments have suggested themselves.

Text analysis experiments that go to questions of transmission and causation would be very beneficial, in particular, since they should scale nicely with the size of collections. However, experiments with human subjects might, in the end, be the most

successful approach to this problem. Human subject experiments could test processes of selection, transmission, copying, and the causal link between a linguistic trait and its subsequent reappearance. Quite a range of such experiments immediately suggest themselves. For instance, by tracking the activities of human authors one could immediately identify cases of convergent analogies versus homologies due to common descent (because, after all, you could ask the authors how they arrived at their wording).

9.3.2 Improvements to the text analysis system

The Microevolutionary Language Theory is supported by an active language replicator model, and empirical support has been acquired thanks to efforts of the CAMEL software system. The active language replicator model proposes that linguistic traits replicate autocatalytically; however, the exact makeup of these traits is the outcome of particular engineering decisions encoded in the CAMEL system. A variety of enhancements to CAMEL should improve the quality of the replicators distilled.

In Section 4.4.5, I described how resolving anaphoric references would improve the SVO lexico-syntactic replicators. These, along with the other replicator forms, would be improved by a variety of other sophisticated text processing steps. For instance, scoping negative particles would allow the system to recognize the difference between “Bill Clinton is guilty” versus “Bill Clinton is not guilty.” Any differential replication of these two replicators would be instructive.

In the previous section, I argued that the Microevolutionary Language Theory would be better supported if more examples of complex design were accumulated. The text analysis system could be enhanced to discover language replicators that have undergone certain lexical semantic innovations and these, then, could be inspected for complex adaptive design. In ongoing work, I have developed a system that tracks the context vectors for lexical replicators and, via a polylogarithmic model, is able to discern lexical traits undergoing significant semantic shifts or expansions. Further work along these lines is required before I will have convincing results.

A number of other language replicators would be worthy of study. For instance, other phrasal groups might undergo active replication (e.g., why not prepositional phrases?). While such co-occurring terms provide useful units of study, other statistical approaches to collocation would be worth trying (along with refinements to the LSI approach described in Section 4.3). For example, techniques based on mea-

asures of mutual information might distill collocatives that are undergoing active replication.

One general problem of the CAMEL system is that it is quite specific to the English language. This is due both to a reliance on an English tagger and parser as well as to a presumption of many of the properties of English (e.g., the SVO ordering, prominence of verbs, etc.). Studies across languages would be very instructive but would require significant modifications to the system.

Some of these shortcomings and extensions are discussed in Gatherer & McEwan (1998) with responses in Best & Pocklington (1999).

9.3.3 New studies and environments

This dissertation should, if it is to be judged a success, suggest a wide range of further experiments.

Certainly, a variety of ecological studies are possible such as the one described in Chapter 8. How do replicators and text interact across a range of environments (e.g., across newsgroups, media, genre, time, people, politics)? Can we find a variety of ecological interactions? Can we perform experiments by injecting engineered texts into an environment and following their progress?

These sorts of experiments might answer some of our questions on macroevolution. Environmentally minded observations such as these, taken over sufficient time periods, might support the observation (retrospectively) of speciation events and related large-scale evolutionary phenomena.

Towards these aims in particular, valuable future work could develop the CAMEL system into a turnkey investigators benchtop. The system might then be used by a variety of researchers who focus it on their text collections of choice. Such a system should include a suite of visualization modules that would allow researchers to observe (perhaps in realtime) the dynamics of text and replicators.

9.4 Summary

In this dissertation I have proposed a Microevolutionary Language Theory which states that complex functional design accumulates at simple levels within human natural language, due to the evolutionary algorithm. I have reported on a collection of experiments which support this theory. These experiments have relied on a soft-

ware system which distills active language replicators from collections of texts.

These active replicators are discovered by:

- identifying linguistic traits that reoccur over time within a set of texts,
- arguing that relevant subsets of these texts describe an evolutionary lineage because they share copied traits,
- demonstrating that the appearance of some of these reoccurring traits correlates with a measure of survivability for these texts.

In the first two chapters of this dissertation I illustrated the concept of an active replicator with two examples: the beak of finches from the Galápagos archipelago and phrases, such as, “right wing ignorance,” from posts to the alt.politics.clinton newsgroup. In Chapter 3, I described the CAMEL software system which was used to identify active language replicators within collections of text. In Chapter 4, I reported on active replicators at a variety of linguistic levels: lexical, lexical co-occurrence, lexico-syntactic, and syntactic. In Chapter 5, I worked on a collection of theoretical problems including framing the Microevolutionary Language Theory within current models of Universal Darwinism. In Chapter 6 and 7, I relayed some outcomes of the theory and model by reporting on studies of competition between linguistic replicators and on improvements in text retrieval. And, in the penultimate chapter, I discussed related work.

In this chapter, I’ve concluded with an attempt at answers to three questions: What is it good for? How is it different? and, Where can we go from here? Like all enthusiastic students, I hope to have managed only partial and tentative answers to these questions — indeed, I would love, in the end, to be surprised at how this theory, model, and system is used, who it excites, and what it illuminates.

Conclusions

Appendix A

The precision/recall experiments of Chapter 7 employed the following twelve queries, here listed in TREC-3 long format (Voorhees & Harman, 1996):

Query 1

<title> Airplane Crashes and other Aviation Accidents

<desc> A relevant document will refer to a plane crash, airliner crash, airplane crash or other aviation accident or air disaster. It may discuss investigations of the cause of the accident conducted by the Federal Aviation Administration or the National Transportation Safety Board. These investigations may concern the aviation industry as a whole or may deal with elements of the planes such as the flight recorder or voice recorder.

Query 2

<title> Bill Clinton and his relationship with campaign finance controversies

<desc> A relevant document will refer to campaign finance controversies involving President Bill Clinton. They could relate to taking foreign donations, campaign finance reform bills in Congress, or controversial donations to the Democratic National Committee (DNC). Documents could also address issues such as illegal soft money contributions or a plot by the Chinese to influence American elections.

Query 3

<title> Congress and its reaction to immigration

<desc> A relevant document will refer to US immigration and the response of the United States Congress. Documents will possibly deal with illegal aliens, illegal immigrants, or refugees and political asylum. Documents could discuss new immigration laws, the Immigration and Naturalization Service (INS), immigration officials and the Justice Department.

Appendix A

Query 4

<title> The AIDS virus

<desc> Relevant documents will discuss the AIDS virus, HIV. Documents could mention modes of viral transmission, such as unprotected sexual contact or needle sharing among intravenous drug users. Documents could also mention treatments for the infection including AZT, protease inhibitors, and combination drug therapy. Mention could also be made of medical research aimed at treating the illness.

Query 5

<title> Scandal in the US military

<desc> A relevant document will refer to any sex scandal or alleged incidents of sexual misconduct within the US military. Documents may mention sexual harassment and discrimination, as well as rape accusations, accusations of indecent assault, sexual assault or sexual abuse. Incidents mentioned could also involve the breaking of military law through acts of consensual sex. These may often involve military officers who have committed adultery.

Query 6

<title> Oklahoma City Bombing

<desc> Relevant documents will discuss the bombing, by Timothy McVeigh and Terry Nichols, of the Alfred P. Murrah Federal Building in Oklahoma City which took place on April 19, 1995 and killed 168 people. Documents could make mention of the Oklahoma City bombing trial and the involvement of federal prosecutors in the case. Mention could also be made of the death penalty and prosecutors seeking a death sentence for the Oklahoma City bombers

Query 7

<title> Tobacco Lawsuits

<desc> Relevant documents will discuss the tobacco lawsuits filed by the state attorneys general against the tobacco companies. Issues related to this are the search for tobacco industry documents on smoking related illnesses, the desire of the cigarette makers to have protection from future lawsuits, the national settlement

with cigarette companies, and the filing of class action lawsuits against the cigarette companies by sick smokers.

Query 8

<title> Genetic Research

<desc> Relevant documents will discuss genetic research. Articles could discuss human cloning or animal cloning and the societal implications of such actions. Articles could also discuss the human genome project and its impact on medical research and pharmaceutical research. Topics in medicine, such as gene replacement therapy and genetic testing, would also be relevant. The use of genetically engineered vegetables in food, or the use of genetic engineering in agriculture, would indicate relevant documents.

Query 9

<title> Cancer

<desc> A relevant document will make reference to any cancer related illnesses. Documents could mention various types of cancer, such as breast cancer and lung cancer. Documents could also make reference to cancer research and medical advances in cancer treatment as well as journals where these discoveries are publicized, such as the New England Journal of Medicine and the Journal of the American Medical Association.

Query 10

<title> Sexual scandals of Bill Clinton

<desc> A relevant document will refer to the scandals involving sexual relations between President Bill Clinton and other women such as Monica Lewinsky and Paula Jones. Documents could make mention of the independent counsel investigations of his conduct and the investigator Kenneth Starr. Documents could also discuss the impeachment proceedings in the House of Representatives and discussions of congressional censure.

Query 11

<title> Space exploration and research

<desc> Relevant documents will make mention of any effort made by world governments or groups to explore and conduct research in or about outer space. Docu-

Appendix A

ments may make mention of existing space stations, such as the Russian Mir, and the current effort to build the International Space Station. Documents could also mention scientists and astronomers conducting research on the planets and galaxies of the universe.

Query 12

<title> The conflict in the former Yugoslavia

<desc> Relevant documents will detail the recent conflicts in the former Yugoslavia. Documents could mention the ethnic groups of Yugoslavia, including Croatians, Serbians, and Bosnians. The alleged ethnic cleansing, genocide and war crimes committed by the Serbs, Slobodan Milosevic and Radovan Karadzic could be mentioned, as well as the war in Kosovo and the displacement of ethnic Albanian refugees.

References

- Aarts, J., & Meijs, W. (Ed.) (1990). *Theory and practice in corpus linguistics*. Amsterdam: Rodopi.
- Ackley, D.H., & Littman, M.L. (1994). Altruism in the evolution of communication. In R.A. Books & P. Maes (Eds.), *Artificial Life IV* (pp. 40-48). Cambridge, MA: MIT Press.
- Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998, February). Topic detection and tracking pilot study final report. In *Proceedings of the Broadcast News Transcription and Understanding Workshop* (Sponsored by DARPA).
- Anttila, R. (1989). *Historical and comparative linguistics* (2nd ed.). Amsterdam: John Benjamins Publishing Company.
- Arlotto, A. (1972). *Introduction to historical linguistics*. Washington, D.C.: University Press of America, Inc.
- Arita, T., & Taylor, C.E. (1996). A simple model for the evolution of communication. In L.J. Fogel, P.J. Angeline & T. Bäck (Eds.), *Proceedings of the Fifth Annual Conference on Evolutionary Programming* (pp. 405-409). Cambridge, MA: MIT Press.
- Asher, N., & Lascarides, A. (1995). Lexical disambiguation in a discourse context. *Journal of Semantics*, 12(1), 69-108.
- Asher, N., & Wada, H. (1988). A computational account of syntactic, semantic and discourse principles for anaphora resolution. *Journal of Semantics*, 6, 309-344.
- Aunger, R. (1999). A report on the conference 'Do Memes Account for Culture?' held at King's College, Cambridge. *Journal of Memetics* [Online] 3(2). Available http://www.cpm.mmu.ac.uk/1999/vol3/cambridge_conference.html.
- Aitchison, J. (1998). On discontinuing the continuity-discontinuity debate. In J.R. Hurford, M. Studdert-Kennedy & C. Knight (Eds.), *Approaches to the evolution of language* (pp. 17-29). Cambridge, UK: Cambridge University Press.
- Ayala, F.J. (1983). Microevolution and macroevolution. In D.S. Bendall (Ed.), *Evolution from molecules to men*. Cambridge: Cambridge University Press.

References

- Baayen, H. (1993). Statistical models for word frequency distributions: A linguistic evaluation. *Computers and the Humanities*, 26, 347-363.
- Baayen, R.H., & Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language*, 72(1), 69-96.
- Back, T. (1996). *Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*. New York: Oxford University Press.
- Bak, P., & Sneppen, K. (1993). Punctuated equilibrium and criticality in a simple model of evolution. *Physical Review Letters*, 71(24), 4083-4086.
- Baker, C.L. (1996). *English syntax* (2nd ed.). Cambridge, MA: MIT Press.
- Baker, M.C. (1996). Depauperate meme pool of vocal signals in an island population of singing honeyeaters. *Animal Behavior*, 51, 853-858.
- Baldwin, B. (1995). *CogNIAC: A discourse processing engine*. Unpublished doctoral dissertation, University of Pennsylvania.
- Baldwin, J.M. (1996). A new factor in evolution. In R.K. Belew & M. Mitchell (Eds.), *Adaptive individuals in evolving populations: Models and algorithms* (pp. 59-80). Reading, MA: Addison-Wesley. (Original work published in 1896.)
- Barkow, J.H. (1989). *Darwin, sex, and status: Biological approaches to mind and culture*. Toronto: University of Toronto Press.
- Barkow, J.H., Cosmides, L., & Tooby, J. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford: Oxford University Press.
- Barnbrook, G. (1996). *Language and computers: A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University press.
- Batali, J. (1994). Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax. In R. Brooks & P. Maes (Eds.), *Artificial Life IV* (pp. 160-171). Cambridge, MA: MIT Press.
- Bates, E., & MacWhinney, B. (1990). Welcome to functionalism. *Behavioral and Brain Sciences*, 13, 727-728.
- Bauer, L. (1994). *Watching English change: An introduction to the study of linguistic change in standard Englishes in the Twentieth Century*. London: Longman.

-
- Bedau, M.A. (1995). Three illustrations of artificial life's working hypothesis. In W. Banzhaf & F.H. Eeckman (Eds.), *Evolution and biocomputation: Computational models of evolution* (pp. 53-68). Berlin: Springer.
- Bedau, M.A., & Packard, N.H. (1992). Measurement of evolutionary activity, teleology, and life. In C.G. Langton, C.E. Taylor, J.D. Farmer & S. Rasmussen (Eds.), *Artificial Life II* (pp. 431-461). Redwood City, CA: Addison-Wesley.
- Bedau, M.A., Snyder, E., Brown, C.T., & Packard, N.H. (1997). A comparison of evolutionary activity in artificial evolving systems and in the biosphere. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life, ECAL97* (pp. 125-134). Cambridge, MA: MIT Press/Bradford Books.
- Below, R.K. (1990). Evolution, learning, and culture: Computational metaphors for adaptive algorithms. *Complex Systems*, 4, 11-49.
- Below, R.K. (2000). *Finding out about: A cognitive perspective on search engine technology and the WWW*. Cambridge, UK: Cambridge University Press.
- Berry, M.W. (1992). Large-scale sparse singular value computations. *The International Journal of Supercomputer Applications*, 6(1), 13-48.
- Berry, M., Do, T., O'Brien, G., Krishna, V., & Varadhan, S. (1993). *SVDPACKC (Version 1.0) user's guide*. (Tech. Rep. CS-93-194). University of Tennessee, Knoxville, Computer Science Department.
- Berry, M.W., & Fierro, R.D. (1995). *Low-rank orthogonal decompositions for information retrieval applications*. (Tech. Report CS-95-284). University of Tennessee, Knoxville, Computer Science Department.
- Best, M.L. (1997). Models for interacting populations of memes: Competition and niche behavior. *Journal of Memetics* [On-line serial] 1(2). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Best, M.L. (1998a). An ecology of text: Using text retrieval to study alife on the net. *Journal of Artificial Life*, 3(4): 261-287.
- Best, M. L. (1998b). Memes and genetic opposition. In *Proceedings of the 15th International Congress on Cybernetics*. Namur, Belgium: Association Internat. de Cybernetique.
- Best, M.L. (1998c). Memes on memes: A critique of memetic models. *Journal of Memetics* [On-line serial], 2(1). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.

References

- Best, M.L. (1999a). *Adaptive value within natural language discourse*. Manuscript submitted for publication.
- Best, M.L. (1999b). Coevolving mutualists guide simulated evolution. In W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela & R.E. Smith (Eds.) *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 99)* (p. 941). San Francisco, CA: Morgan Kaufmann Publishers.
- Best, M.L. (in press). How culture can guide evolution: An inquiry into gene/meme enhancement. *Journal of Adaptive Behavior*.
- Best, M.L. & Pocklington, R. (1999). Meaning as use: Transmission fidelity and evolution in NetNews. *Journal of Theoretical Biology*, 196, 389-395.
- Betzig, L. (Ed.) (1997). *Human nature: A critical reader*. Oxford: Oxford University Press.
- Biber, D. (1993). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26, 331-345.
- Bichakjian, B.H. (1995). Essentialism in language: A convenient, but fallacious premise. In S. Puppel (Ed.), *The biology of language* (pp. 33-59). Amsterdam: John Benjamins Publishing Co.
- Blackmore, S. (1998). Imitation and the definition of a meme. *Journal of Memetics* [Online] 2(2). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Blackmore, S. (1999). *The meme machine*. Oxford: Oxford University Press.
- Boag, P.T., & Grant, P.R. (1984). The classical case of character release: Darwin's Finches (*Geospiza*) on Isla Daphne Major, Galápagos. *Biological Journal of the Linnean Society*, 22, 243-287.
- Bock, W.J. (1980). The definition and recognition of biological adaptation. *American Zoologist*, 20, 217-227.
- Bookstein, A., Klein, S.T., & Raita, T. (1995). Detecting content-bearing words by serial clustering — Extended abstract. In *Proceedings of the 18th ACM-SIGIR Conference* (pp. 319-327). New York: ACM Press.
- Boyd, R., & Richerson, P.H. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.

-
- Breden, F., & Hausfater, G. (1990). Selection within and between social groups for infanticide. *American Naturalist*, 136, 637-688.
- Breden, F., & Wade, M. (1989). Selection within and between kin groups of the imported willow leaf beetle. *American Naturalist*, 134, 35-50.
- Brockman, J. (1995). *The third culture*. New York: Touchstone.
- Brodie, R. (1996). *Virus of the mind: The new science of the meme*. Seattle: Integral Press.
- Brown, C.H. (1979). A theory of lexical change (with examples from folk biology, human anatomical partonomy and other domains). *Anthropological Linguistics*, 21(6), 257-276.
- Bura, S. (1994). MINIMEME: Of life and death in the Noosphere. In D. Cliff, P. Husbands, J-A Meyer & S.W. Wilson (Eds.), *From animals to animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior* (pp 479-486). Cambridge, MA: MIT Press.
- Calvin, W.H. (1997). The six essentials? Minimal requirements for the Darwinian bootstrapping of quality. *Journal of Memetics* [On-line serial] 1(1). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Campbell, D. (1989). Introduction to nonlinear phenomena. In D. Stein (Ed.), *Complex systems: SFI studies in the sciences of complexity* (pp. 3-105). Redwood City, CA: Addison-Wesley Publishing Company.
- Campbell, D.T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review*, 67, 380-400.
- Campbell, D.T. (1965). Variation and selective retention in sociocultural evolution. In H.R. Barringer, G.I. Blanksten & R.W. Mack (Eds.) *Social change in developing areas: A reinterpretation of evolutionary theory*. Cambridge, MA: Schenkman Publishing Company.
- Campbell, D.T. (1974). Evolutionary epistemology. In P.A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 413-463). La Salle, IL: Open Court.
- Caplan, A.L. (Ed.) (1978). *The sociobiology debate: Readings on ethical and scientific issues*. New York: Harper & Row, Publishers.
- Cariani, P. (1991). Emergence and artificial life. In C.G. Langton, C. Taylor, H.D. Farmer & S. Rasmussen (Eds.), *Artificial life II: Santa Fe Institute studies in the sciences of complexity, volume X* (pp. 775-797). Redwood City, CA: Addison-

References

Wesley.

- Cavalli-Sforza, L., & Feldman, M. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton, NH: Princeton University Press.
- Cavnan, W.B. (1975). Using an N-gram-based document representation with a vector processing retrieval model. In D.K. Harman (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST Special Publication 500-226). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Chatfield, C. (1989). *The analysis of time series: An introduction*. London: Chapman and Hall.
- Chen, H. (1995). Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science*, 46(3), 194-216.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Churchill, W.S. (1943). *The end of the beginning: War speeches*. Boston: Little, Brown and Company.
- Clarke, D.D., & Nerlich, B. (1991). Word-waves: A computational model of lexical semantic change. *Language & Communication*, 11(3), 227-238.
- Clutton-Brock, T.H., & Harvey, P.H. (1979). Comparison and adaptation. *Proceedings of the Royal Society of London B*, 205, 547-565.
- Colless, D.H. (1985). On "character" and related terms. *Systematic Zoology*, 34, 229-233.
- Conexor oy (1998a). *The EngCG-2 tagger*. [Online]. Available: <http://www.conexor.fi/engcg2.html#1> [July 31, 1999].
- Conexor oy (1998b). *Functional dependency grammar of English (FDG)*. Available: <http://www.conexor.fi/fdg.html#1> [July 31, 1999].
- Cosmides, L., & Tooby, J. (1989). Evolutionary psychology and the generation of culture, Part II: Case study: A computational theory of social exchange. *Ethology and Sociobiology*, 10, 51-97.
- Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J.H. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163-228). New York: Oxford

University Press.

Cosmides, L., Tooby, J., & Barkow, J.H. (1992). Introduction: Evolutionary psychology and conceptual integration. In J.H. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 3-15). New York: Oxford University Press.

Crow, J.F., & Kimura, M. (1970). *An introduction to population genetics theory*. New York: Harper and Row.

Darmesteter, Arsène (1886). *The life of words as the symbols of ideas*. London: Kegan Paul, French & Co.

Darwin, C. (1964). *On the origin of species*. Cambridge, MA: Harvard University Press. (Original work published 1859.)

Dawkins, R. (1976). *The selfish gene*. Oxford, UK: Oxford University Press.

Dawkins, R. (1978). Replicator selection and the extended phenotype. *Zeitschrift für Tierpsychologie*, 47, 61-76.

Dawkins, R. (1982). *The extended phenotype*. San Francisco: WH Freeman.

Dawkins, R. (1983). Universal Darwinism. In D.S. Bendall (Ed.), *Evolution from molecules to man* (pp. 403-425). Cambridge, UK: Cambridge University Press.

Dawkins, R. (1993). Viruses of the mind. In G. Stocker & C. Schöpf (Eds.), *Dennett and his critics: Demystifying mind* (pp. 40-47). Berlin: Springer-Verlag.

Deacon, T.W. (1997). *The symbolic species: The co-evolution of language and the brain*. New York: W.W. Norton & Company.

Deb, D. (1996). Of cast net and caste identity: Memetic differentiation between two fishing communities of Karnataka. *Human Ecology*, 24(1), 109-123.

de Boer, B. (1997). Generating vowel systems in a population of agents. In P. Husbands & I. Harvey (Eds.), *Fourth European Conference on Artificial Life* (pp. 503-510), Cambridge, MA: MIT Press.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

Dennett, D.C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.

References

- Dornhaus, A., & Chittka, L. (1999). Evolutionary origins of bee dances. *Nature*, 401, 38.
- Dumais, S.T. (1992). LSI meets TREC: A status report. In D. Harman (Ed.), *The First Text REtrieval Conference (TREC-1)* (NIST Special Publication 500-207, pp. 137-152). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Dumais, S.T. (1993). Latent Semantic Indexing (LSI) and TREC-2. In D. Harman (Ed.), *The Second Text REtrieval Conference (TREC-2)* (NIST Special Publication 500-215, pp. 219-230). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Dunbar, R. (1996). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.
- Durham, W.H. (1990). Advances in evolutionary culture theory. *Annual Review of Anthropology*, 19, 187-210.
- Durham, W.H. (1991). *Coevolution: Genes culture and human diversity*. Stanford, CA: Stanford University Press.
- Edmonds, B. (1998). On modeling in memetics. *Journal of Memetics* [On-line serial] 2(2). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Eigen, M. (1992). *Steps towards life: A perspective on evolution*. Oxford, UK: Oxford University Press.
- Eldredge, N. (1989). *Macro-evolutionary dynamics: Species, niches, and adaptive peaks*. New York: McGraw-Hill Publishing Company.
- Encyclopædia Britannica Online*. (1999). Socrates [Online]. Available: http://search.eb.com/bol/topic?artcl=109554&seq_nbr=1&page=p&isctn=6 [1999, August 24].
- Endler, J.A. (1986). *Natural selection in the wild*. Princeton, NJ: Princeton University Press
- Fagan, J.L. (1987). *Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods*. Unpublished doctoral dissertation, Cornell University.
- Feldman, G. (1998). *Birds of the Galápagos* [Online]. Available: <http://www.fas.harvard.edu/~gfeldman/birds.html> [1999, September 21].

-
- Feldman, M.W., & Laland, K.N. (1996). Gene-culture coevolutionary theory. *TREE*, 11, 453-457.
- Feresten, S. (1995). The soup nazi (A. Ackerman, Director). In *Seinfeld*. New York: NBC.
- Fisher, R.A. (1912). *Social selection*. Paper presented at the meeting of the Cambridge University Eugenics Society, Cambridge, UK.
- Flyvbjerg, H., Bak, P., Jensen, M.H., & Sneppen, K. (1995). A self-organized critical model for evolution. In E. Mosekilde & O.G. Mouritsen (Eds.), *Modeling the dynamics of biological systems: Nonlinear phenomena and pattern formation* (pp. 269-288). Berlin: Springer.
- Foltz, P.W. (1990). Using Latent Semantic Indexing for information filtering. In *Proceedings of the 5th Conference on Office Information Systems*. ACM SIGOIS Bulletin 11(2-3), 40-47.
- Forrest, S. (1990). Emergent computation: Self-organizing, collective, and cooperative phenomena in natural and artificial computing networks. *Physica D*, 42, 1-11.
- Fox, C. (1992). Lexical analysis and stoplists. In W.B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 102-130). Upper Saddle River, NJ: Prentice Hall.
- Frakes, W.B. (1992a). Introduction to information storage and retrieval systems. In W.B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 1-12). Upper Saddle River, NJ: Prentice Hall.
- Frakes, W.B. (1992b). Stemming algorithms. In W.B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 131-160). Upper Saddle River, NJ: Prentice Hall.
- Frakes, W.B., & Baeza-Yates, R. (Eds.) (1992). *Information retrieval: Data structures & algorithms*. Upper Saddle River, NJ: Prentice Hall.
- Francis, W. N., & Kucera, H. (Eds.) (1979). *Manual of information to accompany a Standard Corpus of Present-Day Edited American English for use with digital computers* (rev. ed.). Providence, RI: Brown University, Department of Linguistics.
- Francis, W.N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

References

- Fristrup, K. (1992). Character: Current usages. In E.F. Keller & E.A. Lloyd (Eds), *Keywords in evolutionary biology* (pp. 45-51). Cambridge, MA: Harvard University Press.
- Furnas, G.W., Deerwester, S., Dumais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A., & Lochbaum, K.E. (1988). Information retrieval using a Singular Value Decomposition model of latent semantic structure. In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval (SIGIR)*. New York: Association for Computing Machinery.
- Gabora, L.M. (1995). Meme and variations: A computational model of cultural evolution. In L. Nadel & D.L. Stein (Eds.), *1993 Lectures in complex systems*. Reading, MA: Addison-Wesley.
- Gale, W.A., Church, K.W., & Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26, 415-439.
- Galef, B.G., Jr. (1994). Olfactory communications about foods among rats: A review of recent findings. In B.G. Galef, Jr., M. Mainardi & P. Valsecchi (Eds.), *Behavioral aspects of feeding* (pp. 83-102). Chur, Switzerland: Harwood Academic Publishers.
- Gatherer, D. (1998). Why the 'Thought Contagion' metaphor is retarding the progress of memetics. *Journal of Memetics* [On-line serial] 2(2). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Gatherer, D., & McEwan, N.E. (1998). On units of selection in cultural evolution. *Journal of Theoretical Biology*, 192, 409-413.
- Gerard, R.W., Kluckhohn, C., & Rapoport, A. (1956). Biological and cultural evolution: Some analogies and explorations. *Behavioral Science*, 1, 6-34.
- Gerbig, A. (1997). *Lexical and grammatical variation in a corpus: A computer-assisted study of discourse on the environment*. Frankfurt: Peter Lang.
- Ghiselin, M.T. (1997). *Metaphysics and the origin of species*. Albany, NY: State University of New York Press.
- Giraldeau, L-C, & Lefebvre, L. (1987). Scrounging prevents cultural transmission of food-finding behavior in pigeons. *Animal Behavior*, 35, 387-394.
- Goldberg, D.E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley Publishing Company.
- Gottman, J.M. (1981). *Time-series analysis: A comprehensive introduction for*

social scientists. Cambridge, UK: Cambridge University Press.

- Gould, S.J. (1980). Is a new and general theory of evolution emerging? *Paleobiology*, 6, 119-130.
- Gould, S.J. (1984). Covariance sets and ordered geographic variation in *Cerion* from Aruba, Bonaire and Curacao: A way of studying nonadaptation. *Systematic Zoology*, 33(2), 217-237.
- Gould, S.J. (1991). *Bully for brontosaurus*. New York: Norton.
- Gould, S.J. (1997a). Darwinian fundamentalism. *The New York Review of Books*, 44(10), 34-37.
- Gould, S.J. (1997b). Evolution: The pleasures of pluralism. *The New York Review of Books*, 44(11), 47-52.
- Gould, S.J., & Eldredge, N. (1977). Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology*, 3, 115-151.
- Gould, S.J., & Lewontin, R.C. (1979). The spandrels of San Marco and the Panglossian paradigm. *Proceedings of the Royal Society of London, B*, 205, 581-598.
- Gould, S.J., Raup, D.M., & Sepkoski, Jr., J.J. (1977). The shape of evolution: A comparison of real and random clades. *Paleobiology*, 3, 23-40.
- Gould, S.J., & Vrba, E.S. (1982). Exaptation - a missing term in the science of form. *Paleobiology*, 8, 4-15.
- Grant, P.R. (1985). Interspecific competition in fluctuating environments. In J.M. Diamond & T.J. Case (Eds.), *Community Ecology* (pp. 172-191). New York: Harper and Row.
- Grant, P.R. (1986). *Ecology and evolution of Darwin's Finches*. Princeton, NJ: Princeton University Press.
- Green, D.G. (1993). Emergent behavior in biological systems. In D.G. Green & T. Bossomaier (Eds.), *Complex systems: From biology to computation* (pp. 24-35). Amsterdam: IOS Press.
- Greenberg, J.H. (1992). Preliminaries to a systematic comparison between biological and linguistic evolution. In J.A. Hawkins & M. Gell-Mann (Eds.), *The evolution of human languages* (pp. 139-158). Redwood City, CA: Addison-Wesley.

References

- Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Boston: Kluwer Academic Publishers.
- de Grolier, E. (Ed.) (1983). *Glossogenetics: The origin and evolution of language*. Chur, Switzerland: Harwood Academic Publishers.
- Gruhl, D. (2000). *The search for meaning in large text databases*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.
- Györi, G. (1995). Animal communication and human language: Searching for their evolutionary relationship. In S. Puppel (Ed.) *The biology of language* (pp. 99-126). Amsterdam: John Benjamins Publishing Co.
- Hales, D. (1998). Selfish memes & selfless agents - Altruism in the swap shop. In *Proceedings of the 3rd International Conference on Multi-Agent Systems*. Los Gatos, CA: IEEE Press.
- Hamilton, W.D. (1964). The genetical evolution of social behavior. I and II. *Journal of Theoretical Biology*, 7, 1-52.
- Harman, D. (1992). Ranking algorithms. In W.B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 363-392). Upper Saddle River, NJ: Prentice Hall.
- Harman, D.K., & Voorhees, E.H. (Eds.) (1997). *Proceedings of the fifth Text REtrieval Conference (TREC-5)* (NIST Special Publication 500-238). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Hashimoto, T. (1994). Usage-based structuralization of relationships between words. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life, ECAL97* (pp. 483-492). Cambridge, MA: MIT Press/Bradford Books.
- Hawkins, J.A., & Gell-Mann, M. (Eds.) (1992). *The evolution of human languages*. Redwood City, CA: Addison-Wesley Publishing Company.
- Heikkilä, J. (1995a). ENGTWOL English lexicon: Solutions and problems. In F. Karlsson, A. Voutilainen, J. Heikkilä & A. Anttila (Eds.), *Constraint grammar: A language-independent system for parsing unrestricted text* (pp. 133-163). Berlin: Mouton de Gruyter.
- Heikkilä, J. (1995b). A TWOL-based lexicon and feature system for English. In F. Karlsson, A. Voutilainen, J. Heikkilä & A. Anttila (Eds.), *Constraint grammar: A language-independent system for parsing unrestricted text* (pp. 103-131).

Berlin: Mouton de Gruyter.

- Heyes, C.M., & Dawson, G.R. (1990). A demonstration of observational learning using a bidirectional control. *Quarterly Journal of Experimental Psychology*, *42B*, 59-71.
- Heyes, C.M., & Galef, B.G., Jr. (Eds.) (1996). *Social learning in animals: The roots of culture*. San Diego, CA: Academic Press.
- Hildebrand-Nilshon, M. (1995). From proto-language to grammar: Psychological considerations for the emergence of grammar in language evolution. In S. Puppel (Ed.), *The biology of language* (pp. 127-145). Amsterdam: John Benjamins Publishing Co.
- Hobbs, J.R. (1978). Resolving pronoun references. *Lingua*, *44*, 311-338.
- Holland, J.H. (1992). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence* (Rev. ed.). Cambridge, MA: MIT Press.
- Horton, M., & Adams, R. (1987). *Standard for interchange of USENET messages*. Internet RFC-1036.
- Howell, D.C. (1995). *Fundamental statistics for the behavioral sciences* (3rd ed.). Belmont, CA: Duxbury Press.
- Hoyle, F., & Wickramasinghe, N.C. (1981). *Evolution from space*. London: J.M. Dent.
- Huffman, M.A. (1996). Acquisition of innovative cultural behaviors in nonhuman primates: A case study of stone handling, a socially transmitted behavior in Japanese Macques. In C.M. Heyes & B.G. Galef, Jr. (Eds.), *Social learning in animals: The roots of culture* (pp. 267-290). San Diego, CA: Academic Press.
- Hull, D. (1980). Individuality and selection. *Annual Review of Ecology and Systematics*, *11*, 311-332.
- Hull, D. (1982). The naked meme. In H.C. Plotkin (Ed.), *Learning, development, and culture: Essays in evolutionary epistemology* (pp. 273-325). Chichester: John Wiley & Sons.
- Hull, D. (1988). *Science as a Process*. Chicago: University of Chicago Press.
- Hull, D. (1992). Individual. In E.F. Keller & E.A. Lloyd (Eds.), *Keywords in evolutionary biology* (pp. 180-187). Cambridge, MA: Harvard University Press.

References

- Hull, D. (1999a). Strategies in meme theory. *Journal of Memetics* [On-line serial], 3. Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Hull, D. (1999b, June). *Taking memetics seriously: Memetics will be what we make it*. Paper presented at the meeting Do Memes account for Culture, Cambridge, UK.
- Hurford, J.R. (1990). Beyond the roadblock in linguistic evolution studies. *Behavioral and Brain Sciences*, 13(4), 736-737.
- Hurford, J.R. (1991). An approach to the phylogeny of the language faculty. In J.A. Hawkins & M. Gell-Mann (Eds.), *The evolution of human languages* (pp. 273-303). Redwood City, CA: Addison-Wesley Publishing Company.
- Hurford, J.R. (1992). The evolution of the critical period for language acquisition. *Cognition*, 40, 159-201.
- Hurford, J.R., Studdert-Kennedy, M., & Knight, C. (Eds.) (1998). *Approaches to the evolution of language*. Cambridge, UK: Cambridge University Press.
- Jackendoff, R.S. (1972). *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.
- Jain, A.K., & Dubes, R.C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.
- James, W. (1880). Great men, great thoughts, and the environment. *The Atlantic Monthly*, 46(276), 441-459.
- Jane, E. A., & Lampert, M.D. (Eds.) (1993). *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Erlbaum.
- Jeffers, R.H., & Lehiste, I. (1979). *Principles and methods for historical linguistics*. Cambridge, MA: MIT Press.
- Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English for use with digital computers*. Oslo, Norway: Department of English, University of Oslo.
- Jones, W.P., & Furnas, G.W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6), 420-442.
- de Jong, M. (1999). Survival of the institutionally fittest concepts. *Journal of Memetics* [On-line serial] 3(1). Available: [http://www.cpm.mmu.ac.uk/jom-](http://www.cpm.mmu.ac.uk/jom-emit)

emit.

- Kantor, B., & Lapsley, P. (1986). *Network news transfer protocol: A proposed standard for the stream-based transmission of news*. Internet RFC-977.
- Karahalios, K. (1998). *Loom* [Online]. Available: <http://www.media.mit.edu/~kkarahal/loom/> [October 26, 1998].
- Karllsson, F., Voutilained, A., Heikkilä, J., & Anttila, A. (Eds.). (1995). *Constraint grammar: A language-independent system for parsing unrestricted text*. Berlin: Mouton de Gruyter.
- Karttunen, L. (1983). KIMMO: A two-level morphological analyzer. *Texas Linguistic Forum*, 22, 217-228.
- Keller, R. (1994). *On language change: The invisible hand in language*. London: Routledge.
- Kettridge, R., & Lehrberger, J. (Eds.) (1982). *Sublanguage: Studies of language in restricted semantic domains*. Berlin: Walter de Gruyter.
- Kirby, S. (in press). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models*. New York: Cambridge University Press.
- Kjellmer, G. (1994). Lexical differentiators of style: Experiments in lexical variability. In U. Fries, G. Tottie & P. Schneider (Eds.), *Creating and using English language corpora* (pp. 117-126). Amsterdam: Rodopi.
- Kleparski, G. (1986). *Semantic change and componential analysis; An inquiry into pejorative developments in English*. Regensburg: Friedrich Pustet.
- Koskenniemi, K. (1983). Two-level morphology: A general computational model for word-form production and generation. (Pub. No. 13). Helsinki, Finland: University of Helsinki, Department of General Linguistics.
- Kucera, H. (1992). Brown corpus. In S. C. Shapiro (Ed.), *Encyclopedia of artificial intelligence* (Vol. 1, pp. 128-130). New York: John Wiley & Sons.
- Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Kuper, A. (1999, June). *If memes are the answer, what is the question*. Paper presented at the meeting Do Memes account for Culture, Cambridge, UK.
- Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of

References

- Pennsylvania Press.
- Lachlan, R.F., & Slater, P.J.B. (1999). The maintenance of vocal learning by gene-culture interaction: The cultural trap hypothesis. *Proceedings of the Royal Society of London, B*, 266, 701-706.
- Laland, K.N., Odling-Smee, J., & Feldman, M.W. (in press). Niche construction, biological evolution and cultural change. *Behavioral and Brain Sciences*.
- Laland, K.N., Kumm, J., & Feldman, M.W. (1995) Gene-culture coevolutionary theory: A test case. *Current Anthropology*, 36(1), 131-156.
- Lappin, S., & Leass, H.J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), 535-561.
- Lass, R. (1997). *Historical linguistics and language change*. Cambridge, UK: Cambridge University Press.
- Lenormand, T., Bourguet, D., Guillemaud, T., & Raymond, M. (1999). Tracking the evolution of insecticide resistance in the mosquito *Culex pipiens*. *Nature*, 400, 861-864.
- Lewis, D.D., & Croft, W.B. (1990). Term clustering of syntactic phrases. In J.-L. Vidick (Ed.), *Proceedings of the Thirteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 385-404). New York: ACM Press.
- Lewis, D.D., & Sparck Jones, K. (1996). Natural language processing for information retrieval. *Communications of the ACM*, 39(1), 92-101.
- Lewontin, R.C. (1970). The units of selection. *Annual Review of Ecological Systems*, 1, 1-18.
- Lewontin, R.C. (1978). Adaptation. *Scientific American*, 239(3), 212-230.
- Lieberman, P. (1992). On the evolution of human language. In J.A. Hawkins & M. Gell-Mann (Eds.), *The evolution of human languages* (pp. 21-47). Redwood City, CA: Addison-Wesley Publishing Company.
- Lieberman, P. (1998). *Eve spoke: Human language and human evolution*. New York: W.W. Norton & Company.
- Liiv, H. (1997). A method for singling out representative linguistic features. *Journal of Quantitative Linguistics*, 4(1-3), 139-142.
- Lloyd, E.A. (1992). Unit of selection. In E.F. Keller & E.A. Lloyd (Eds), *Keywords*

-
- in evolutionary biology* (pp. 334-340). Cambridge, MA: Harvard University Press.
- Locke, J.L., & Hauser, M.D. (1999). Sex and status effects on primate volubility: Clues to the origin of vocal languages. *Evolution and Human Behavior*, 20, 151-158.
- Love, M. (1980). The alien strategy. *Natural History*, 89(5), 30-32.
- Lumsden, C.H., & Wilson, E.O. (1981). *Genes, mind and culture: The coevolutionary process*. Cambridge, MA: Harvard University Press.
- Lyell, C. (1863). *The geological evidences of the antiquity of man, with remarks on theories of the origin of species by variation*. London: John Murray.
- Lynch, A. (1996). *Thought contagion: How belief spreads through society*. New York: Basic Books.
- Lynch, A. (1998). Units, events and dynamics in memetic evolution. *Journal of Memetics* [On-line serial] 2(1). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Maas, U. (1994). 'Nazi-Journalismus' und ähnliche komposita: Zu den Spuren der Vergangenheitsbewältigung im Wortschatz. *Grazer-linguistische-monographien*, 11, 129-143.
- MacLennan, B. (1991). Synthetic ethology: An approach to the study of communication. In C.G. Langton, C. Taylor, H.D. Farmer, & S. Rasmussen (Eds.), *Artificial life II: Santa Fe Institute studies in the sciences of complexity, volume X*. Redwood City, CA: Addison-Wesley.
- MacWhinney, B. (Ed.) (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Mair, C. (1997). Parallel corpora: A real-time approach to the study of language change in progress. In M. Ljung (Ed.), *Corpus-based studies in English* (pp. 195-209). Amsterdam: Rodopi.
- Marr, A. (1998, August 16). This dictionary describes a language eaten into, a mongrel language changing shape at great speed. *The Guardian*, p. 21.
- Marr, D. (1982). *Vision: A computational investigation into human representation and processing of visual information*. San Francisco: Freeman.
- Marsden, P. (1998). Operationalising memetics: Suicide, the Werther Effect, and the

References

- work of David P. Phillips. In *Proceedings of the 15th International Congress on Cybernetics (Association Internat. de Cybernetique, Namur)*.
- Marsden, P. (1999). [Review of the book *Thought contagion: How believe spreads through society*]. *Journal of Artificial Societies and Social Simulation* [On-line serial], 2. Available: <http://www.soc.surrey.ac.uk/JASSS/JASSS.html>.
- Martinet, A. (1975). *Studies in functional syntax*. Munich, Germany: W. Fink. (Original work published 1960).
- Matthews, P. (1997). *The concise Oxford dictionary of linguistics*. Oxford, UK: Oxford University Press.
- May, R.M. (1981). Models for two interacting populations. In R.M. May (Ed.), *Theoretical ecology: Principles and applications* (pp. 78-104). Oxford, UK: Blackwell Scientific Publications.
- Maynard Smith, J. (1972). *On evolution*. Edinburgh, UK: Edinburgh University Press.
- Maynard Smith, J. (1976). Group selection. *Quarterly Review of Biology*, 51, 277-283.
- Mayr, E. (1991). *One long argument: Charles Darwin and the genesis of modern evolutionary thought*. Cambridge, MA: Harvard University Press.
- McCleary, R., & Hay, R.A. (1980). *Applied time series analysis for the social sciences*. Beverly Hills, CA: Sage Publications.
- McMahon, A.M.S. (1994). *Understanding language change*. Cambridge, UK: Cambridge University Press.
- Mel'cuk, I. A. (1988). *Dependency syntax: Theory and practice*. Albany, NY: State University Press of New York.
- Menczer, F., & Belew, R.K. (2000). Adaptive retrieval agents: Internalizing local context and scaling up to the web. *Machine Learning*, 29, 2/3.
- Messer, D.J. (1995). Language acquisition and the essentialist-evolutionist debate. In S. Puppel (Ed.), *The biology of language* (pp. 173-194). Amsterdam: John Benjamins Publishing Co.
- Metzenberg, R.L. (1990). The role of similarity and difference in fungal mating. *Genetics*, 125, 457-462.
- Milic, L.T. (1966). Unconscious ordering in the prose of Swift. In Leed, J. (Ed.), *The*

-
- computer and literary style*. Kent, OH: Kent State University Press.
- Miller, G.A. (Ed.) (1990). *Journal of Lexicography*, 3(4).
- Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA: MIT Press.
- Moritz, E. (1990). Replicator based knowledge representation and spread dynamics. In *Proceedings of the 1990 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 256-259). Los Gatos, CA: IEEE Press.
- Moukas, A., & Maes, P. (1998). Amalthea: An evolving multi-agent information filtering and discovery system for the WWW. *Autonomous Agents and Multi-Agent Systems*, 1, 59-88.
- Nerlich, B. (1990). *Change in language: Whitney, Bréal, and Wegener*. London: Routledge.
- Niyogi, P., & Berwick, R.C. (1996). A language learning model for finite parameter spaces. *Cognition*, 61, 161-193.
- Noble, J., & Cliff, D. (1996). On simulating the evolution of communication. In P. Maes, M.J. Mataric, J. Meyer, J. Pollack & S. Wilson (Eds.), *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 608-617). Cambridge, MA: MIT Press.
- Noble, W., & Davidson, I. (1996). *Human evolution, language and mind: A psychological and archaeological inquiry*. Cambridge, UK: Cambridge University Press.
- Nowak, M.A., & Krakauer, D.C. (1999). The evolution of language. *Proceedings of the National Academy of Science*, 96, 8028-8033.
- Oakes, M.P. (1998). *Statistics for corpus linguistics*. Edinburgh, UK: Edinburgh University Press.
- Oostdijk, N. (1988). A corpus linguistic approach to linguistic variation. *Literary and Linguistic Computing*, 3(1), 12-25.
- Paley, W. (1803). *Natural theology: Or evidences of the existence and attributes of the deity, collected from the appearances of nature* (5th ed.). London: Faulder.
- Percival, R.S. (1994). Dawkins and incurable mind viruses? Memes, rationality and evolution. *Journal of Social and Evolutionary Systems*, 17(3), 243-286.
- Piaget, J. (1980). *Adaptation and intelligence: Organic selection and phenocopy*.

References

- Chicago: University of Chicago Press.
- Pianka, E.R. (1981). Competition and niche theory. In R.M. May (Ed.), *Theoretical ecology: Principles and applications* (pp. 167-196). Oxford, UK: Blackwell Scientific Publications.
- Pielou, E.C. (1969). *An introduction to mathematical ecology*. New York: Wiley-Interscience.
- Pinker, S. (1994). *The language instinct*. New York: HarperPerennial.
- Pinker, S. (1997). *How the mind works*. New York: W.W. Norton & Company.
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13, 707-784.
- Plotkin, H. (1994). *Darwin machines and the nature of knowledge*. Cambridge, MA: Harvard University Press.
- Pocklington, R., & Best, M.L. (1997). Cultural evolution and units of selection in replicating text. *Journal of Theoretical Biology*, 188, 79-87.
- Pocklington, R., & Best, M.L. (1999). *Numerical taxonomy as an exploratory tool in high dimensional semantic spaces*. Unpublished manuscript.
- Polomé, E.C. (Ed.) (1990). *Research guide on language change*. Berlin: Mouton de Gruyter.
- Popper, K. (1965). *Conjectures and refutations: The growth of scientific knowledge*. New York: Harper and Row.
- Popper, K. (1972). *Objective knowledge: An evolutionary approach*. Oxford, UK: Oxford University Press.
- Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.
- Puppel, S. (Ed.) (1995). *The biology of language*. Amsterdam: John Benjamins Publishing Co.
- Pustejovsky, J. Bergler, S., & Anick, P. (1993). Lexical semantic techniques for corpus analysis. In S. Armstrong (Ed.), *Using large corpora* (pp. 291-318). Cambridge, MA: MIT Press.
- Radford, A. (1997). *Syntax: A minimalist introduction*. Cambridge, UK: Cambridge University Press.

-
- Rasmussen, E. (1992). Clustering algorithms. In W.B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 419-442). Upper Saddle River, NJ: Prentice Hall.
- Raup, D.M., & Gould, S.J. (1974). Stochastic simulation and evolution of morphology — Towards a nomothetic paleontology. *Systematic Zoology*, 23, 305-322.
- Reeve, H.K., & Sherman, P.W. (1993). Adaptation and the goals of evolutionary research. *Quarterly Review of Biology*, 68(1), 1-32.
- Rhodes, B., & Starner, T. (1996). Remembrance agent: A continuously running automated information retrieval system. In *Proceedings of Practical Applications of Intelligent Agents and Multi-Agent Technology (PAAM)* (pp. 487-495). Blackpool, UK: Practical Application Company.
- Richards, W. (Ed.) (1988). *Natural computation*. Cambridge, MA: MIT Press.
- Robinson, P.M.W., & O'Hara, R.J. (1996). Cladistic analysis of an Old Norse manuscript tradition. In *Research in humanities computing 4: Selected papers from the ALLC/ACH Conference* (pp. 115-137). Oxford, UK: Clarendon Press.
- Rocha, M. (1997). A probabilistic approach to anaphora resolution in dialogues in English. In M. Ljung (Ed.). *Corpus-based studies in English* (pp. 261-279). Amsterdam: Rodopi.
- Rose, N. (1998). Controversies in meme theory. *Journal of Memetics* [On-line serial], 2(1). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Ruhlen, M. (1992). An overview of genetic classification. In J.A. Hawkins & M. Gell-Mann (Eds.), *The evolution of human languages*. Redwood City, CA: Addison-Wesley Publishing Company.
- Sack, W. (1994). On the computation of point of view. In *Proceedings of the National Conference of Artificial Intelligence (AAAI 94)*. Menlo Park, CA: AAAI Press.
- Sack, W. (1999, June). Diagrams of social cohesion. Demonstration presented at the meeting of the Association for Computational Linguistics (ACL'99), University of Maryland, College Park, MD.
- Sack, W. (in press, a). Conversation map: A content-based Usenet newsgroup browser. In D. Riecken, D. Benyon & H. Lieberman (Eds.), *Proceedings of the International Conference on Intelligent User Interfaces*. New York: Association for Computing Machinery.

References

- Sack, W. (in press, b). Discourse diagrams: Interface design for very large-scale conversations. In T. Erikson & S. Herring (Eds.), *Proceedings of the 33rd Hawaii International Conference on System Sciences, Persistent Conversations Track*. New York: Association for Computing Machinery.
- Salmons, J. (1990). The context of language change. In E.C. Polomé (Ed.), *Research guide on language change*. Berlin: Mouton de Gruyter.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.
- Salton, G., & McGill, M. (1983). *An introduction to modern information retrieval*. New York: McGraw-Hill.
- Samuelsson, C., & Voutilainen, A. (1997). Comparing a linguistic and a stochastic tagger. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann.
- Saunders, G.M., & Pollack, J.B. (1996). The evolution of communication schemes over continuous channels. In P. Maes, M.J. Mataric, J. Meyer, J. Pollack & S. Wilson (Eds.), *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 580-589). Cambridge, MA: MIT Press.
- Schmidt-Nielsen, K. (1983). *Animal physiology* (3rd ed.). Cambridge, UK: Cambridge University Press.
- Schütze, H., & Pedersen, J.O. (1995, April). Information retrieval based on word senses. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV.
- Semon, R.W. (1921). *The mneme*. (L. Simon, Trans.). London: George Allen & Unwin Ltd. (Original work published 1908.)
- Sereno, M.I. (1991). Four analogies between biological and cultural/linguistic evolution. *Journal of Theoretical Biology*, 151, 467-507.
- Sheth, B., & Maes, P. (1993). Evolving agents for personalized information filtering. In *Proceedings of the Ninth Conference on Artificial Intelligence for Applications* (pp. 345-352). Los Alamitos, CA: IEEE Computer Society Press.
- Simpson, G.G. (1944). *Tempo and mode in evolution*. New York: Columbia University Press.

-
- Simpson, G.G. (1970). Uniformitarianism: An inquiry into principle, theory, and method in geohistory and biohistory. In M.K. Hecht & W.C. Steere (Eds.), *Essays in evolution and genetics in honor of Theodosius Dobzhansky* (pp. 43-96). New York: Appleton-Century-Crofts.
- Simpson, J.A., & Weiner, E.S.C. (Prepared). (1989). *The Oxford English dictionary* (2nd Ed.). Oxford, UK: Oxford University Press.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.
- Smadja, F. (1993). Retrieving collocations from text: Xtract. In S. Armstrong (Ed.), *Using large corpora* (pp. 143-177), Cambridge, MA: MIT Press.
- Smeaton, A.F., & van Rijsbergen, C.J. (1988). Experiments on incorporating syntactic processing of user queries into a document retrieval strategy. In *Eleventh International Conference on Research & Development in Information Retrieval* (pp. 31-51). New York: ACM Press.
- Smillie, D. (1995). Biological and cultural factors in the evolution of language. In S. Puppel (Ed.), *The biology of language* (pp. 266-276). Amsterdam: John Benjamins Publishing Co.
- Sober, E. (1984). *The nature of selection: Evolutionary theory in philosophical focus*. Cambridge, MA: MIT Press.
- Spark Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworths.
- Spark Jones, K., & Bates, R.G. (1977). *Research on automatic indexing 1974-1976* (2 volumes). (Technical report). Cambridge, UK: University of Cambridge, Computer Laboratory.
- Sparck Jones, K., & van Rijsbergen, C. (1975). *Report on the need for and provision of an "ideal" information retrieval test collection*. (British Library Research and Development Report 5266). Cambridge, UK: University of Cambridge, Computer Laboratory.
- Speel, H-C. (1997). A memetic analysis of policy making. *Journal of Memetics* [Online serial] 1(2). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Spartus, E. (1997). Smokey: Automatic flame recognition. In *Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence (IAAI-97)* (pp. 1058 - 1065). Menlo Park, CA: AAAI Press.

References

- Steels, L. (1996). Emergent adaptive lexicons. In P. Maes, M.J. Mataric, J. Meyer, J. Pollack & S. Wilson (Eds.), *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 562-567). Cambridge, MA: MIT Press.
- Steels, L. (1998a). The origins of ontologies and communication conventions in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 1, 169-194.
- Steels, L. (1998b). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103, 1-24.
- Stevick, R.D. (1963). The biological model and historical linguistics. *Language*, 39, 159-169.
- Strzalkowski, T., & Carballo, J.P. (1994). Recent developments in natural language text retrieval. In D.K. Harman (Ed.), *The Second Text REtrieval Conference (TREC-2)* (NIST Special Publication 500-215, pp. 123-136). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Strzalkowski, T., & Carballo, J.P. (1996). Natural language information retrieval: TREC-4 report. In D.K. Harman (Ed.), *The Fourth Text REtrieval Conference (TREC-4)* (NIST Special Publication 500-236, pp. 245-257). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Strzalkowski, T., Carballo, J.P., & Marinescu, M. (1995). Natural language information retrieval: TREC-3 report. In D.K. Harman (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST Special Publication 500-226, pp. 39-54). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Strazalkowski, T., Guthrie, L., Karlgren, J., Leistensnider, J., Lin, F., Perez-Carballo, J., Straszheim, T., Wang, J., & Wilding, J. (1997). Natural language information retrieval: TREC-5 report. In D.K. Harman & E.M. Voorhees (Eds.), *Proceedings of the fifth Text REtrieval Conference (TREC-5)* (NIST Special Publication 500-238, pp. 291-314). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Strazalkowski, T., Lin, F., & Perez-Carballo, J. (1998). Natural language information retrieval TREC-6 report. In E.M. Voorhees & D.K. Harman (Eds.), *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (NIST Special Publication 500-240, pp. 347-366). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Tapanainen, P., & Järvinen, T. (1997, March). A non-projective dependency parser.

-
- In *5th Conference on Applied Natural Language Processing*. Washington D.C.
- Tooby, J., & Cosmides, L. (1989). Evolutionary psychology and the generation of culture, Part I: Theoretical considerations. *Ethology and Sociobiology*, *10*, 29-49.
- Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J.H. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford, UK: Oxford University Press.
- Times of London* (1989). Reprinted in J.A. Simpson, & E.S.C. Weiner (Prepared). *The Oxford English dictionary* (2nd ed.). Oxford: Oxford University Press (Original work published 1930, September 19).
- Tracy, L. (1996). Genes, memes, templates, and replicators. *Behavioral Science*, *41*, 205-214.
- Traugott, E.C. (1985). On regularity in semantic change. *Journal of Literary Semantics*, *14*(3), 155-173.
- Turkel, W.J. (1997). What can we learn from computational models of language evolution? Unpublished manuscript.
- Voorhees, E.M. (1998). Using WordNet for text retrieval. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 285-303). Cambridge, MA: MIT Press.
- Voorhees, E.M., & Harman, D. (1997). Overview of the Fifth Text REtrieval Conference (TREC-5). In D.K. Harman & E.M. Voorhees (Eds.), *Proceedings of the fifth Text REtrieval Conference (TREC-5)* (NIST Special Publication 500-238, pp. 1-28). Washington, D.C.: Department of Commerce, National Institute of Standards and Technology.
- Vose, M.D., & Liepins, G.E. (1991). Punctuated equilibria in genetic search. *Complex Systems*, *5*, 31-44.
- Voutilainen, A. (1995a). Experiments with heuristics. In F. Karlsson, A. Voutilainen, J. Heikkilä & A. Anttila (Eds.), *Constraint grammar: A language-independent system for parsing unrestricted text* (pp. 294-314). Berlin: Mouton de Gruyter.
- Voutilainen, A. (1995b). Morphological disambiguation. In F. Karlsson, A. Voutilainen, J. Heikkilä & A. Anttila (Eds.), *Constraint grammar: A language-independent system for parsing unrestricted text* (pp. 165-283). Berlin: Mouton

References

- de Gruyter.
- Waddington, C.H. (1957). Selection by, of, and for. In C.H. Waddington (Ed.), *The Strategy of the genes* (pp 59-108). London: Allen and Unwin.
- Watson, J. (1976). *Molecular biology of the gene*. Menlo Park, CA: Benjamin Cummings.
- Weasel, Z. (1997, November 4). Re: PUBLIC DEBATE CHALLENGE TO 'ZEPP WEASEL' (Repeated). In alt.politics.usa.constitution [Online]. Available Usenet News [1997, November 4].
- Wei, W.W.S. (1990). *Time series analysis: Univariate and multivariate methods*. Redwood City, CA: Addison-Wesley Publishing Company, Inc.
- Weiner, J. (1995). *The beak of the finch*. New York: Vintage Books.
- Weinreich, U., Labov, W., & Herzog, M.I. (1968). Empirical foundations for a theory of language change. In W.O. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics: A symposium* (pp. 95-195). Austin, TX: University of Texas Press.
- Werner, G.M., & Dyer, M.G. (1991). Evolution of communication in artificial organisms. In C.G. Langton, C. Taylor, H.D. Farmer & S. Rasmussen (Eds.), *Artificial Life II: Santa Fe Institute studies in the sciences of complexity, volume X*. Redwood City, CA: Addison-Wesley.
- Whiten, A., & Custance, D. (1996). Studies of imitation in chimpanzees and children. In C.M. Heyes & B.G. Galef, Jr. (Eds.), *Social learning in animals: The roots of culture* (pp. 291-318). San Diego, CA: Academic Press.
- Whiten, A., Goodall, J., McGrew, W.C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C.E.G., Wrangham, R.W., & Boesch, C. (1999). Cultures in chimpanzees. *Nature*, 399, 682-685.
- Wickler, W. (1976). Evolution-oriented ethology, kin selection, and altruistic parasites. *Zeitschrift für Tierpsychologie*, 42, 206-214.
- Wilkins, D.P. (1996). Natural tendencies of semantic change and the search for cognates. In M. Durie & M. Ross (Eds.), *The comparative method reviewed: Regularity and irregularity in language change*. Oxford, UK: Oxford University Press.
- Wilkins, J.S. (1998). What's in a meme? Reflections from the perspective of the history and philosophy of evolutionary biology. *Journal of Memetics* [On-line]

-
- serial] 2(1). Available: <http://www.cpm.mmu.ac.uk/jom-emit>.
- Williams, G.C. (1966). *Adaptation and natural selection*. Princeton, NJ: Princeton University Press.
- Williams, G.C. (1992). *Natural Selection*. Oxford, UK: Oxford University Press.
- Wilson, E.O. (1978). *On human nature*. Cambridge, MA: Harvard University Press.
- Worden, R.P. (1998, April). *Words, memes and language evolution*. Paper presented at the meeting Evolution of Language, London.
- Wright, R. (1994). *The moral animal: Evolutionary psychology and everyday life*. New York: Vintage Books.
- Wright, S. (1982). Character change, speciation, and the higher taxa. *Evolution*, 36, 427-433.
- Wynne-Edwards, V.C. (1962). *Animal dispersion in relation to social behavior*. Edinburgh, UK: Oliver & Boyd.
- Xiong, R. (1998). *Netviz* [Online]. Available: <http://netscan2.sscnet.ucla.edu/netviz/> [October 26, 1998].
- Yang, J-J, & Korfhage, R.R. (1993). Query optimization in information retrieval using genetic algorithms. In S. Forrest (Ed.), *Proceedings of the Fifth International Conference on Genetic Algorithms* (pp. 603-611). San Mateo, CA: Morgan Kaufmann Publishers.
- Zbigniew, M. (1992). *Genetic algorithms + data structures = evolution programs*. Berlin: Springer-Verlag.
- Zernik, U. (Ed.) (1991). *Lexical acquisition: Exploiting on-line resources to build a lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.