Human Information Retrieval

Julian Warner

Human Information Retrieval

History and Foundations of Information Science
Edited by Michael Buckland and Jonathan Furner

Human Information Retrieval, by Julian Warner

Human Information Retrieval

Julian Warner

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Sabon by The MIT Press.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Warner, Julian, 1955–
Human information retrieval / Julian Warner.
p. cm.—(History and foundations of information science)
Includes bibliographical references and index.
ISBN 978-0-262-01344-4 (alk. paper)
1. Information retrieval. I. Title.
ZA3075.W375 2010
025.04—dc22

2009010120

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments

The John Campbell Trust supported a presentation on the labor theoretic approach to information retrieval at the Annual Meeting of the American Society for Information Science and Technology, Providence, RI, November 2004. I received hospitality and assistance from the Research Centre for the Social Sciences (RCSS), University of Edinburgh, where most of the articles from which material has been derived were originally written during study leave from the Queen's University of Belfast, February–July 2005.

I am indebted to the unnumbered students at the Queen's University of Belfast and Indiana University who interactively contributed to the development of the research themes expounded by taking courses in Information Policy, Communicating Electronically, and Computer Mediated Communication, 2001–2007.

Michael Buckland, Elisabeth Davenport, and Madeleine Coyle also gave me support and encouragement.

Chapters 2–4 are adapted, respectively, from the following series of articles:

Warner, J. 2007. Selection power and selection labor for information retrieval. *Journal of the American Society for Information Science and Technology* 58 (7): 915–923.

Warner, J. 2007. Description and search labor for information retrieval. *Journal of the American Society for Information Science and Technology*. 58 (12): 1783–1790.

Warner, J. 2008. A labor theoretic approach to information retrieval. *Journal of the American Society for Information Science and Technology* 59 (5): 731–741.

Chapter 2 also incorporates material from the following article:

Warner, J. 2000. In the catalogue ye go for men: evaluation criteria for information retrieval systems. *Aslib Proceedings* 52 (2): 76–82.

Chapters 5–8 incorporate material from the following articles:

Warner, J. 2007. Analogies between linguistics and information theory. *Journal of the American Society for Information Science and Technology* 58 (3): 309–321.

Warner, J. 2007. Linguistics and information theory: analytic advantages. *Journal of the American Society for Information Science and Technology* 58 (2): 275–285.

They also include some material published in the following book chapter.

Warner, J. 2008. Materializing communication concepts: linearity and surface in linguistics and information theory. In *Exploration of space, technology, and spatiality: Interdisciplinary perspectives*, ed. P. Turner, S. Turner, and E. Davenport., 196–213. Hershey: Information Science Reference Press.

1
Introduction

Information retrieval is of high contemporary significance, diffusing into ordinary discourse and everyday practice. Recently information retrieval has changed rapidly, particularly through the influence of Internet search engines. Practical understanding of how to use systems has been in advance of fuller theoretic understanding, particularly when concerned with transformations of meaning. The breadth of the area—the variety of disciplines that have and potentially could contribute—may have inhibited the development of understanding. A deep divide remains between centrally relevant activities and their associated cultures, between library and information science and Internet search engines. The limited theoretical development implies a need for a deeper—more comprehensive and inclusive—understanding of information retrieval, which should reveal structure and underlying patterns in a complex, differentially understood, and apparently chaotic area. Ideally, a deeper understanding or theory should be congruent with practical understanding and everyday practice. It should also be explicitly articulated, yielding knowledge which can be returned to the real world to inform deliberate intervention in system design and use rather than unconsciously reproducing patterns of activity.

The overall intention is to develop a labor theoretic approach to information retrieval. The immediate concern in this chapter involves the initial components for this approach. This chapter also reviews existing evaluative traditions and indicates the possibility for synthesis within a labor theoretic approach. Chapter 2 introduces and develops crucial concepts of selection power and selection labor, both as concepts and activities in themselves and for the relation between them.

Labor, choice, and technology are fundamental to human experience. In the Judeo-Christian tradition, once out of Eden we are condemned to labor and compelled to choose. Technology may have been noticed less explicitly, but it is equally pervasive in post-Edenic experience, both as agrarian and industrial—or *productive*—and information technologies. Physical and mental labor are usually considered separately from each other, but acknowledgment of the mental components of physical labor and physical elements in mental labor have moved recently toward synthesis, and an emerging view of intelligence concerns "quality of our bodies as much as our minds" (Gosden 2003, 31–33, 119). Mental or informational labor has been recognized as both an independent activity and an adjunct to obtaining physical control over the environment (Webster 2002, 15). Types of mental labor have been differentiated, with semantic labor distinguished from syntactic labor (Warner 2005a). Classically from Aristotle, choice, or deliberation, is the product of mental labor. Late-twentieth-century developments in information technology constitute a revolution in the mechanization of mental labor (Minsky 1967, 2), embodied in the computer as a universal information machine that developed from mid- and late-nineteenth-century antecedents in special-purpose information machines. Both productive and information technologies are human constructions—the products of human physical and intellectual labor drawing upon natural resources and preexisting human constructions (Warner 2004, 5–35). An understanding of information retrieval constructed from labor, choice, and technology promises to be deeply rooted in human experience and to offer a radical depth of understanding.

Power in explanation can be demonstrated by the ability to absorb elements from previous models as special cases of the new model, indicative of the history of a true science, while discarding those elements that have obstructed understanding.

Existing Models

Information science has developed existing evaluative models that should be absorbed into the new model and offer some elements for synthesis and advancement, with diffusion into computer science, librarianship, and indexing. Some discussions of the information society are concerned with

informational or mental labor, and this offers a more indirect resource that can also be absorbed.

Information and Computer Science

The dominant tradition for evaluating information retrieval systems in information science emerged nearly simultaneously and partly independently in the United States and in the United Kingdom during the early- to mid-1950s (Ellis 1996, 1–22) and has since diffused to and been partly absorbed by computer science. The techniques developed for selection and ordering of references and documents have served as exemplars, or demonstrations of possibility, for the increasingly dominant Internet search engines, with some elements of more direct transfer or inheritance in search algorithms derived from information retrieval research (Ellis and Vasconcelos 1999, 8). There were also parallel developments in commercial search services, largely independent in concept but drawing on common technologies, which also served as exemplars and demonstrated commercial feasibility and technical possibility.

The information and computer science tradition has not always been explicit about its own values or examined its own assumptions; without full notice, it has sometimes departed pragmatically from its initial assumptions. It can, however, be broadly characterized as query transformation, with the query articulated verbally in advance of searching and then transformed by a system into a set of records (Heine 1980). The central value of query transformation can be summarized as a system that should deliver all and, insofar as possible, only all records relevant to a given query. Retrieved records are assessed for their relevance to the query and before generating measures of performance, including precision and recall. The adopted methodology induces a bias toward fixing and possibly reifying relevance, reducing it from a concept to a relation between query and document (Ellis 1984, 28–29; 1992; 1996, 11–20). Bibliographic systems, rather than full text, have been the dominant—although not exclusive—subject of study. Humanly assigned indexing has tended to be received as a given, leaving the rationale for such indexes—and their associated labor and costs—unexplored. There has also been an implicit teleology that aims for a perfect system. Evaluation has become an end in itself, sometimes obstructing understanding (Ellis 1984; 1992).

The relation between information technology and the research tradition in information and computer science could be characterized as repression, wherein the repressed reemerges but not at a fully conscious level. Repression is discernible in the insistence that the retrieval processes created and studied are independent of their particular technological instantiation while simultaneously allowing procedures to be strongly determined by contemporary technological possibilities. For instance, the stress on query transformation corresponded to the batch processing embodied in the technology of the 1950s. The theoretical legacy of query transformation has proved difficult to adapt to modern systems, which do not necessarily demand a verbally articulated query in advance of searching and which can be—and are—used interactively. Critiques argued that the assumption of the necessity for a verbally articulated query was intratheoretic (Heine 1980) rather than intrinsic to information seeking; this argument has been substantiated by changes in practice enabled partly by subsequent technological developments. Reemergence of the repressed can be found in the late articulation and still-limited acknowledgment of the identity between primitive operations of information retrieval and logic or computation. Analysis has revealed that the potential transformations for information retrieval on written records or descriptions are variations on primitive operations of sorting or partitioning and the transformation of one symbol into another (Buckland and Plaunt 1994). This can be a regarded as a special case of the known potential for reducing mathematical and logical operations on an object language to the writing, erasure, and substitution of symbols (Ramsey 1925/1990, 165–174) and also corresponding to the primitive computational operations (Warner 1994, 102–103). The paradigm of query transformation can be regarded as largely but not entirely exhausted, becoming increasingly distant from the empirical reality of interactive and distributed systems (Ellis and Vasconcelos 1999, 8), exposing its rigidity if the original distinctions are retained, or surviving by ad hoc modifications to its theoretical base and thereby losing relevance in the first direction and internal intellectual coherence in the other.

Two paradigms—the cognitive and the physical—have been distinguished in information retrieval research, but they share the assumption of the value of delivering relevant records (Ellis 1984, 19; Belkin

and Vickery 1985, 114). For the purposes of discussion here, they can be considered as a single heterogeneous paradigm, linked but not united by this common assumption. The value placed on query transformation is dissonant with common practice, where users may prefer to explore an area and may value fully informed exploration. Some dissenting research discussions have been more congruent with practice, advocating exploratory capability—the ability to explore and make discriminations between representations of objects—as the fundamental design principle for information retrieval systems.

We can acknowledge the utility of techniques developed for selection and ordering of references and documents—in both the experimental tradition and commercial practice—and simultaneously recognize that these techniques are derived from known fundamental computational operations. The techniques have been realized in the special-purpose tools and machines used for information retrieval at the beginning of the research tradition in the 1950s and by the programmed universal information machine of the modern computer. We can fully acknowledge technology rather than repress it and still make a distinction between techniques and values in order to preserve and carry forward what may be valuable from information retrieval research in information and computer science.

Librarianship and Indexing

Compared with the research tradition developed in information science and subsequently diffused to computer science, the historical antecedents for understanding information retrieval in librarianship and indexing are far longer but less widely influential today. They have tended to be less explicit about their evaluative criteria and aims for information retrieval systems, and far less concerned with producing measures of effectiveness. In contrast to information and computer science, they have been associated with the technologies of writing and printing and have had a pronounced preference for direct human description of information objects. Although less immediately pronounced, we can discern a similar pattern of repression regarding technology. The need for descriptions less extensive than the documents described, imposed by storage constraints of inscribed media, and for direct human intervention in the creation of these descriptions, connected with the technical characteristics of writing

and printing, have tended to be universalized and treated as if they were independent of their dominant technological realizations (Wilson 2001). The information and computer science tradition probably inherited the assumed need for brief index descriptions directly from existing information products and not from theories that informed the construction of those products (Cleverdon 1962; Cleverdon, Mills, and Keen 1966). A further limitation of library studies involves its focus on training in the use of information retrieval systems, which often concentrates on the level of system commands rather than understanding their value in communication (Roberts 1989). Disturbing evidence suggests that formal information retrieval systems are marginal in communication (particularly scholarly communication), especially in the sense of information, topic, or subject retrieval rather than document identification (known author or title) and supply (Bath University Library 1971; Smithson 1994).

Two valuable elements are carried forward from librarianship and indexing. The first is a partly implicit stress on selection power, conceived as bibliographic control in librarianship (Wilson 1968) and implied by valuing the discriminatory power of index terms in indexing but made fully explicit here. The second is an acknowledgment of the role of direct human intellectual labor in creating selection power, transforming into a fuller understanding distinguished from specific technological constraints and their partly covert influence on theory and practice.

Information Society Discussions

Information society discussions have given some rather limited attention to information retrieval. For instance, Lyotard comments:

> It is reasonable to suppose that the proliferation of information-processing machines is having, and will continue to have, as much an effect on the circulation of learning as did advancements in human circulation (transportation systems) and later, in the circulation of sounds and visual images (the media). (1984, 4)

Other comments remain similarly unfocused, recognizing the significance of information retrieval but not providing full research or intellectual context for its consideration. In particular, some information society discussions have treated technology unsatisfactorily (Webster 2002), possibly due to wariness about being stigmatized as technologically determinist (Wilson 1996a), and there has been a limited understanding of funda-

mental computer operations. An analytically valuable category of informational labor has begun to be distinguished by some writers (Webster 2002, 15); this book will adopt and further differentiate that distinction, acknowledging the possibility of transferring some forms of mental labor to information technology.

Summary

Different elements from the information retrieval tradition developed in information science from librarianship, indexing, and information society discussions will be selected and carried forward. The utility of the techniques developed by information retrieval research—but not the associated value of query transformation—is acknowledged, with the recollection that techniques are variations on primitive computational transformations. Selection power is adopted from librarianship and indexing as the primary value and the role of direct human labor is both substantiated and critiqued. Informational labor transformed into mental labor to incorporate its historical antecedents is derived from information society discussions. Technology is restored, not repressed, and understanding of the types of mental labor transferable to information technology is informed by the distinction between semantic and syntactic mental labor. A synthesis of existing approaches is envisaged, producing a set of concepts and categories that are simultaneously simpler and more powerful than the query transformation of classic information-retrieval research, more explicit and discriminating than librarianship and indexing (particularly regarding the significance and costs of human mental labor), and fuller and more technologically informed than information society discussions.

This book adopts an inclusive understanding of information retrieval systems, developed from common understandings and conveyed by ostensive exemplification rather than restrictive definition. In particular, the common antithesis between experimental and operational systems is dissolved. The real source of contrast between the types of system likely has been different forms of the description process, particularly the experimental preference for machine generation rather than human selection of index terms and for non-Boolean searches for those descriptions. When made explicit, the basis for the distinctions between experimental and

operational information retrieval systems appears theoretically weak. The distinction is being increasingly eroded in practice, with operational systems possibly selecting records or documents by directly Boolean operations but ordering retrieved documents on the basis of other indicators.

If the proposed model is to be regarded as a scientific advance, it must have a dual aspect that comprehends empirical reality and selectively absorbs existing models. Empirical reality should be explained as fully, powerfully, and as parsimoniously as possible. The pervasive presence of labor, choice, and technology in information retrieval practice promises a strong degree of correspondence to empirical reality. Human labor is immediately present as the description labor of cataloging, classification, and database description. Choice has been persistently embodied in practice and, more recently, increasingly recognized theoretically and valued as both selection from retrieval results and the filtering of information. Diffused from the 1950s, modern information technologies, to which aspects of human mental labor can be and increasingly are transferred, are now pervasive in information retrieval.

The reader can now anticipate this book's approach and structure.

The encompassing theories adapted for development and the mode of presentation derives from the human sciences. The author understands human history as a cumulative progression: human labor acting upon the naturally given and humanly modified environment as the primary source for progressive change developed by, but not exclusive to, Marx. Distinctions from Ferdinand de Saussure's *Course in General Linguistics* are adapted for understanding transformations of meaning in full-text retrieval. Material from information theory and computer science is also understood from the perspective of the human sciences. Information theory is utilized to comprehend patterns of occurrence and recurrence of words and phrases. The discussion is informed throughout by an understanding of the computational process derived directly from automata theory, or the theory of computation.

Characteristic of the human sciences, the mode of presentation is primarily discursive and aims to obtain broad intelligibility. Logically expressed schemata are introduced, accompanied by diagrams more often encountered in the formal sciences but intended here to reveal the underlying

rigor and economy of the discursively expressed argument, parallel to the discourse rather than independent points of departure. Methodologically, the book deliberately excludes some complexity, particularly regarding full-text retrieval, to enable analytic concentration on central or inescapable issues; it also excludes simplified assumptions that can be false or distorting.

Each chapter develops cumulatively from the preceding chapters.

Chapter 2: Selection Power and Selection Labor

Selection power—the ability to make informed choices between objects or representations of objects—is argued from a number of perspectives as the primary aim of information retrieval systems. Similar principles for retrieval are adduced from partly independent discourses, such as the value placed upon an index term's discriminatory power in discussions of indexing and the concept of bibliographic control as mastery over written and published records (UNESCO/Library of Congress 1950, 1). Commonly used systems embody facilities to enhance selection power, and the survival of such systems in the market for information services testifies to the perceived utility of selection power (Swanson 1980, 128). Ordinary discourse comments from consumers of such systems also value control and selection. The concept of exploratory capability developed by critiques of the dominant research tradition is further transformed into selection power, providing more precise and generically applicable analysis. Therefore, understandings embodied in some significant scholarly discourses, in practice and in ordinary discourse, precede theoretical articulation, which continues to value query transformation above selection power in the dominant research tradition. The etymology of intelligence (*inter-legere*: to choose between) implies a link between selection power and both deep and ordinary discourse, or widely diffused aspects of human experience (Stevens 1998, 66). Values for information retrieval are brought into accord with processes by replacing query transformation with selection power.

Selection power is conceived as a fundamental concept that is open to elucidation but not further decomposition into more primitive entities. It is understood as a quality of human consciousness that can be assisted or frustrated by the system's capacity for exploration but is not inher-

ent in the system itself. Under certain historical conditions and levels of technological development, selection power is produced by activities such as cataloguing, classification, description of objects for databases, and searching catalogs and databases, all of which can all be comprehended and understood as selection labor. Thus, selection power is not conceived abstractly, apart from real world circumstances, but operates in relation to considerations of human labor (particularly mental labor), the costs of that labor, and the possibility of transferring particular forms of mental labor to information technology, now primarily computational technologies. A fundamental proposition is developed: selection power is produced by selection labor.

Selection labor is characterized as a form of mental labor and theoretical minima are established for a given collection of objects. The separation of selection labor into description and search labor with the *premodern* technologies of writing and printing on paper is noted. Similarly, the author acknowledges that description and search activities reconverge with computer-based, or *modern*, technologies and also acknowledges the possibility of sustaining analytical distinctions between them.

Chapter 3: Description and Search Labor

Selection labor separates historically into description and search labor and can be analytically decomposed. The activities of description and searching are then more fully characterized empirically as components of selection labor. As forms of mental labor, description and search labor participate in the conditions for labor and mental labor. Concepts and distinctions that apply to physical and mental labor are indicated, introducing the necessity of labor for survival, the idea of technology as a human construction, and the possibility of transferring human labor—including mental labor—to technology. Distinctions specific to mental labor, particularly between semantic and syntactic labor, are introduced. The high cost of human mental labor is also indicated.

Exemplified by cataloging, classification, and database description, description labor is more formally understood as the labor involved in transforming objects into searchable descriptions; it includes interpretation. Search labor is understood as the human labor expended in searching systems. Direct human labor has diminished progressively for both

description and search labor, and its syntactic aspects have transferred to technology effectively compelled by the high relative costs of direct human labor compared to machine processes.

Chapter 4: A Labor Theoretic Approach
The labor theoretic approach to information retrieval that informed chapters 2 and 3 is made fully explicit in chapter 4. The labor theoretic approach has qualities usually desired for a theory that couples comprehensiveness with economy— parsimony, power, and final simplicity. Although the focus is on the computational mode, it acknowledges inheritances from orality and literacy, and the theory can also comprehend oral and written modes. The labor theoretic approach absorbs library and information and Internet activities into a common schema within the computational mode; different aspects of the schema become prominent for each set of activities. The schema developed within the theory is economical, explicitly reduced to a short sequence of clauses, and also represented in diagrammatic form that includes elements of iconicity. A very powerful analysis results from making fully explicit the dynamics that are strongly implicit in current information activities. Once grasped, the theory becomes simple.[1]

Chapter 5: Retrieval from Full Text
The labor theoretic approach can account for the existence of full text retrieval, precisely locating significant changes in description and search labor and processes. However, its analytic power is more fully demonstrated in accounting for the existence of changes and precisely specifying their location than by a fuller understanding of those same qualitative changes. As evidenced by the provision and use of phrase searching, practical understanding of transformations of word meaning and the frequency of word sequences has tended to run ahead of theoretical understanding and articulation. Sources acknowledged in retrieval as relevant to understanding word meanings have exposed a misleading concept of language as a nomenclature but have not fully articulated a positive account of the production of meaning in written language. Therefore, fuller and deliberately articulated understanding requires further development. Ideally, further development should remain consistent with the labor theoretic

approach, incorporate existing sources recognized as revealing, and coincide with practical understandings; but it also should develop dialectically from these bases and may draw upon further material.

To obtain deeper insight into inescapable issues of semantics and syntax—the production of meaning from written language and replication and differences in words and words sequence—requires further theoretical contexts. Therefore, the principal sources selected are Saussurean linguistics (for understanding semantics) and information theory (for insight into syntactics understood as patterns of replication and difference in written language), both continually informed by the theory of computation. Selection power is understood to be inescapably produced by human selection labor that modulates over time.

Chapter 6: A Semantics for Retrieval from Full Text
This chapter is concerned with developing a semantics for written language that incorporates crucial distinctions for understanding retrieval from full text. Established and materially rooted categories in Saussurean linguistics, the syntagma and paradigm—the linear sequence of utterance and the network of associations a word acquires outside a particular syntagma—are adapted to a largely unprecedented purpose: an account of the production of meaning from written language. The inheritance of patterns for the production of meaning and the occurrence and recurrence of words and phrases from oral and written language are also acknowledged and placed in dialogic encounter with the possibilities of computation.

Chapter 7: A Syntactics for Retrieval from Full Text
Understood as the occurrence and recurrence of patterns, the syntactics of written language are equally relevant to understanding retrieval from full text. Material elements from information theory—the message and messages for selection from a source that correspond directly to the categories chosen from Saussurean linguistics—are adapted to gain understanding of patterns in written discourse, which then can be directly and humanly exploited in searching. The theory of computation continues to inform the understanding of patterns of occurrence and recurrence, particularly for operations that effectively include cutting of words from a line of writing, consistent with the fundamental computational operations of writing, erasing, and substituting symbols.

Table 1.1
Structure of the book

<table>
<tr><td colspan="5">Approaches to information retrieval</td><td>Chapter 1. Introduction</td><td rowspan="3">Chapter Four. A labor theoretic approach</td><td rowspan="5">Chapter Nine. Conclusion</td><td rowspan="5">Postscript</td></tr>
<tr><td>Selection power</td><td colspan="4">Selection labor</td><td>Chapter 2. Selection power and selection labor</td></tr>
<tr><td></td><td colspan="2">Description labor</td><td colspan="2">Search labor</td></tr>
<tr><td></td><td>Description labor semantic</td><td>Description labor syntactic</td><td>Search labor semantic</td><td>Search labor syntactic</td><td>Chapter 3. Description and search labor</td></tr>
<tr><td></td><td></td><td>Description processes syntactic</td><td>Search labor semantic</td><td rowspan="3"></td><td>Chapter 5. Retrieval from full text</td><td rowspan="3">Chapter Eight. Semantics and syntactics for retrieval from full text</td></tr>
<tr><td></td><td></td><td>Description processes syntactic</td><td>Search labor semantic</td><td>Chapter 6. A semantics for retrieval from full text</td></tr>
<tr><td></td><td></td><td>Description processes syntactic</td><td>Search labor semantic</td><td>Chapter 7. A syntactics for retrieval from full text</td></tr>
</table>

Chapter 8: Semantics and Syntactics for Retrieval from Full Text
This chapter focuses on bringing semantics and syntactics back to the examples of retrieval given in chapter 5 and then testing and demonstrating their analytic advantages. A fuller example of retrieval is also introduced and considered from the developed perspective.

Chapter 9: Conclusion
The conclusion explores the implications of semantics and syntactics developed for preexisting theories of information retrieval, for the labor theoretic approach, and for the practical evolution of Internet search engines.

Postscript
The postscript addresses the changes in what it means to be human that arise from current developments in information retrieval.

A diagram confirms the structure of the book, indicating the topics covered by each chapter in relation to the categories of the labor theoretic approach developed and adopted (see table 1.1). The absence of a category does not mean that the category has been discarded in later chapters, but rather that it either continues to be incorporated into the theory developed without further significant modification, or that it is no longer highly significant in real world practice or has been analytically excluded to enable clarity of attention. For instance, selection power is incorporated into the further theory developed, while semantic description labor and syntactic search labor have diminished in real world practice for retrieval from full text. Accordingly, both are analytically excluded from the later chapters. The diagram offers a guide for placing individual chapters within the overall argument.

Box 1.1
Historical valuing of selection power

Ay, in the catalogue ye go for men;
As hounds, and greyhounds, mongrels, spaniels, curs,
Shoughs, water-rugs, and demi-wolves, are clept
All by the name of dogs: the valu'd file
Distinguishes the swift, the slow, the subtle,
The housekeeper, the hunter, every one
According to the gift which bounteous Nature
Hath in him clos'd; whereby he does receive
Particular addition, from the bill
That writes them all alike;

—Shakespeare. *Macbeth*. c.1606. III.i.91

Macbeth's questioning of the murderers (Shakespeare, 1606/1988, 77–78) indicates the value historically attached to subtlety of distinctions in the language or lexicon of information retrieval systems. In this respect, the passage anticipates the principle formulated in modern discussions of indexing and classification—the value of an index term lies in its discriminatory power—and is consistent with the valuing of selection power (see chapter 2).

Commentaries have glossed "valu'd" as an adjective derived from the noun (value) and not as the participle of the verb (Shakespeare 1606/1988, 77), implying that values attached to objects are analogous to attributes in modern databases and index terms in information retrieval systems. It could also be read as being valued or valuable, giving "[p]articular addition" or added value. A "file" (Shakespeare 1606/1988, 142) is a list or roll, a highly linear technological form.

While regarded as an image of order (Shakespeare 1606/1988, 77), the passage also contains elements of disorder: interactions between the breeds are listed in tension with the hierarchy from "dogs" to breeds of dogs—"shoughs" (shag-haired dogs) lead to "water-rugs" (rough-haired water dogs). The transition from the first-named type ("hounds") to the last ("demi-wolves") represents a move from domestic to semiferal, with a hint of lycanthropy implied by the analogy with types of men. The imposition of hierarchy also associates closely with political tyranny.

2

Selection Power and Selection Labor

Selection Power

Selection power is understood as the human ability to make informed choices between objects or representations of objects. It is adopted here as the primary value or aim for information retrieval systems, in contrast to the stress of query transformation in the experimental research tradition. Selection power is not arbitrarily asserted; its epistemological and ontological status is clarified, its content conveyed through exemplification, and its value supported by the presence of analogous concepts in separate scholarly and ordinary discourses.

Definition and Elucidation

Like other fundamental concepts, selection power may be difficult to define without becoming tautological. The difficulty of definition could suggest the concept's significance. In the classic sense of decomposition into more primitive concepts, definition is deliberately avoided and would be difficult given the fundamental nature of the principle, but the term is still elucidated and a refusal of explication similarly avoided.

Questions regarding the classic practice of definition as decomposing a concept into known entities appear somewhat playfully in literary sources, partly by implication in linguistics and most explicitly in philosophical texts. When Pip is inducted into written literacy in *Great Expectations*, Dickens plays upon the inevitable circularity of definitions:

> "Your sister's a master-mind. A master-mind."
> "What's that?" I asked, in some hope of bringing him to a stand. But Joe was readier with his definition than I had expected, and completely stopped me by arguing circularly, and answering with a fixed look, "Her."

"And I ain't a master-mind." (1861/1946, 50)

The implied association between the perceived need for definitions and written literacy has been more formally stated, with the commonly encountered insistence on preliminary definitions regarded as a product of the cultural transition from orality to literacy (Goody and Watt 1968).

A structuralist perspective in linguistics implies a finally circular understanding of meaning. In *The Course in General Linguistics*, Saussure regarded linguistic signs as obtaining meaning from their negative differences from other signs within a network of signs (1916/1983; Culler 1988, 52). Most radically (and seemingly independently), Wittgenstein admitted the impossibility of defining primitive signs by further decomposition and advocated elucidation rather than definition:

> The meaning of primitive signs can be explained by elucidations. Elucidations are propositions which contain the primitive signs. They can, therefore, only be understood when the meaning of these signs are already known. (1922/1981, § 3.263)

Wittgenstein continues to pursue rigorous logical development regarding the possibilities of combining primitive signs. Therefore, the commitment here to a logical structure and finally formal and discursive presentation does not imply adherence to logical positivism, where empirical propositions must correspond to sense impressions (Ayer 1936/1980). Technically, this book offers elucidations rather than definitions.

The primitive signs to which Wittgenstein refers correspond to atomic facts also distinguished in the *Tractatus Logico-Philosophicus* and "From the existence or non-existence of an atomic fact we cannot infer the existence or non-existence of another" (Wittgenstein 1922/1981, § 2.062). Selection power is understood as an atomic fact or primitive term not amenable to further decomposition, and elucidations rather than definitions are offered. Furthermore, we cannot infer from selection power that other atomic facts exist. The process of elucidation here will begin with an example and then indicate concepts analogous to selection power in independent scholarly and ordinary discourses, implicitly acknowledging that practical understanding of constructing and using information systems has preceded theoretical articulation. The discussion will rise from the concrete to the abstract.

Example

A researcher might wish to distinguish the private individual Samuel Langhorne Clemens from the author Mark Twain. For this purpose, a valuable system would not conflate the individual's two distinguishable aspects but rather allow differentiation. Later, the same researcher might seek information that combines Mark Twain and Samuel Clemens as a single entity. Therefore, an information retrieval system should be capable of both differentiating and linking occurrences of these different names. Originally conceived as fictional in a double sense (Warner 2000), the example does have real historical roots. Collections of copyright proceedings index Twain's copyright disputes under his legal name (Clemens) without providing a link to his pen name (Copyright Decisions 1909). A generic search for Twain and Clemens as a single entity across different sources would have to adapt the terminology already used to search the particular source in use. While index terms can offer discriminations and links between related subjects, selection power is considered characteristic of human consciousness, derivable from but not inhering in semiotic products.

The relation between combined identity (Mark Twain and Samuel Clemens as author *and* private individual) and separate identities (Mark Twain [author] and Samuel Clemens [private individual]) corresponds to the genus: species relation, with public: not public as the differentiating factor. Figure 2.1 illustrates this relationship. The genus: species relation occurs repeatedly in indexing languages (for instance, in thesaural relations between broader and narrower terms and in the relation of indexing terms taken from a controlled vocabulary to the language of discourse, particularly as generic scope contrasted with specificity). For formal logic, the species: genus is analogous to material implication (p *isamemberof* q has similar truth conditions to p *implies* q), although the variables for material implication could denote objects rather than classes (Bell 1937, volume 2, 491). Material implication has been the most productive and most difficult of logical relations (Quine 1937/1953, 84).

Discrimination between Twain (author) and Clemens (private individual) could be obtained by:

- direct serial reading of relevant texts, where the searcher expends labor in reading and discrimination, possibly creating an index;

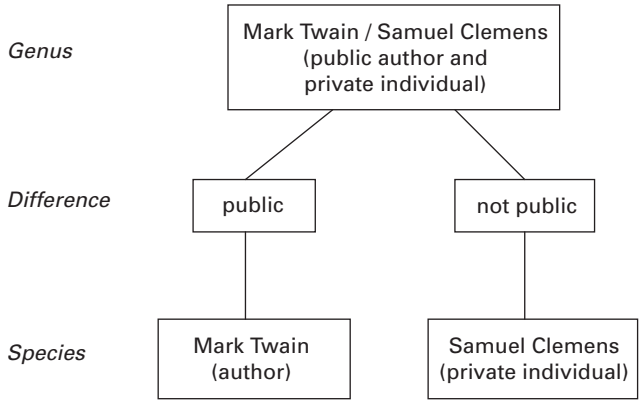


Figure 2.1
Example of a genus-species relation.

• algorithmic and computer-conducted transformations on texts, where the searcher is required to eliminate false recalls and retrieve all relevant instances, both processes complicated by inevitable inconsistencies in the language of discourse; or

• human assignment of index terms and references to sections of discourse, for interpretation by the searcher.

At each stage, the searcher's selection power increases, her/his labor decreases, and the description labor and processes (either transferred to technology or embodied in a human indexer) increase.

Different forms of graphic representation—pictorial, handwritten, or printed—offer different possibilities for algorithmic transformation (see figure 2.2). Curiously, the standard form of computer representation of written language (for instance, ASCII code, which appears more finished and retains less specific production traces than handwriting) is more amenable to algorithmic transformation because the process (keyboarding) is less congealed in the product, as storage in the form of computer memory. Further transformation into more fully graphic representation—rather than directly encoded representation embodied in different file formats—would require a more fully congealed process and would complicate algorithmic transformations.

Thus, selection power operates effectively for an aspect of information retrieval practice often considered separately from other aspects,

Figure 2.2
Contrasting representations of Samuel Clemens / Mark Twain. Source: Railton 2005.

analogous to cataloging in distinction from classification or characterized as data contrasted with subject or topic retrieval. In addition to the value of unified description, recognition of the common principle cataloging shares with classification and subject determination restores theoretical significance and value to it, congruent with the often dominant actual use of information retrieval systems for identifying, recalling, and retrieving known items or the works of a given author (Smithson 1994; Shneiderman 2003, 54).

Scholarly and Ordinary Discourses

Different, and partly independent, scholarly discourses have implicitly endorsed selection power as a design principle for information retrieval systems. Ordinary discourse discussions of information retrieval also value that concept.

The interconnected fields of librarianship and indexing have endorsed as central aims—in different ways and not always explicitly—concepts analogous to or necessary components of selection power. Within librarianship, bibliographic control was seminally defined in the post-1945 period as "mastery over written and published records" (UNESCO/Library of Congress 1950, 1); it is strongly analogous to selection power. Without bibliography, the "records of civilization would be an uncharted chaos of miscellaneous contributions to knowledge, unorganized and inapplicable to human needs" (UNESOC/Library of Congress 1950, viii). At that stage of technological development, the creation of records and indexes for bibliographic organization required direct human intervention

or labor. UNESCO regarded itself as born into "appalling post-war bibliographic chaos" (Murra 1951, 47), and the distribution of responsibility to the national agencies required to produce national bibliographies and allied works on a shared model was conceived as both the remedy for the chaos and the path toward universal bibliographic control. The growth of *WorldCat* (WorldCat 2007) indicates a more recent and less explicitly noticed movement toward universal bibliographic control—particularly regarding monographic literature rather than journal articles—prompted more by internal dynamisms in the process than by imposition and by the possibility of sharing the costs of human description labor by distributing the products of that labor as catalog records. Later sophisticated discussions more directly concerned with selection power also distinguished bibliographic control from bibliographic organization, with organization as the means of control (Wilson 1968), reinforcing the sense of selection power as a property of human consciousnesses enabled by—but not directly inhering in—organization imposed on data.

Indexing values index terms for their discriminatory power, similar to classical logic's concern with differentiae. Discriminatory power could also be regarded as an essential component or organizational factor of enabling discriminatory control or selection power. The technological constraints of written and printed documents, and the need to reduce the cognitive labor of the searcher, compelled briefer and more concise index descriptions than the object described —with some exceptions, such as concordances.

Cybernetics, emerging in its modern form during the immediate post-1945 period concurrently with the formalization of bibliographic control and partly concerned with information technologies envisaged for enhancing bibliographic control, also emphasized control and navigation (Wiener 1954). Cybernetics was understood to embrace the "complex of ideas" represented by

> the study of language … the study of messages as a means of controlling machinery and society, the development of computing machines, and other such automata, certain reflections upon psychology and the nervous system, and a tentative new theory of scientific method. (1954, 15)

In the subsequent development of cybernetics, emphasis on control did not always separate human and machine discrimination. The etymology

of *cybernetics*—given by Wiener, who coined the term in testimony to the Greek *kubernētēs* or *steersman*, understanding the steersman primarily as a human control mechanism (1954, 15), and through its link to Cybernesia (the pilots' festival held in honor of Theseus' navigation to Athens) (Plutarch 100/1914)—points to a deeper level of selection in collective human experience.

A dispute within philosophy (Giambattista Vico's critique of Aristotle) offers a strong and highly significant analogue to exploratory capability. Aristotle's philosophy involved a systematic method of enquiry for classifying an object, in addition to providing a direct and indirect source for subsequent understandings of genus, species, specific difference, synonymy, and equivalence. An enquirer was required to ask a series of questions—Does the thing exist? What is it? How big is it? What is its quality? Vico subjected this method of enquiry to an incisive critique:

> Aristotle's *Categories* and *Topics* are completely useless if one wants to find something new in them. One turns out to be a Llull or Kircher and becomes like a man who knows the alphabet, but cannot arrange the letters to read the great book of nature. But if these tools were considered the indices and ABC's of inquiries about our problem [of certain knowledge] so that we might have it fully surveyed, nothing would be more fertile for research. (1710/1988, 100–101)

The last clause of the critique deserves emphasis: "nothing would be more fertile for research." While retaining some of Aristotle's techniques, Vico avoided his rigidity, transforming it into a systematic and effective means for enhancing knowledge of an object. Analogously, while rejecting a query's rigid transformation into a set of records assumed as desirable in information retrieval research, similar techniques can explore the domain of discourse covered by the information retrieval system. This process of categorization would be somewhat similar to collecting documents under the terms of a humanly assigned metalanguage when describing them, which can then be exploited in searching. Thus, classification schemes (and their analogues in thesaural relations among indexing terms) become valuable exploratory devices.

Dickens's *Hard Times* provides another supporting analogue, although the fiction also enacts a critique of Platonic and Aristotelian philosophy. The logical distinctions exemplified in Bitzer's definition of a horse—"Quadruped. Graminivorous. Forty teeth, namely twenty-four grinders,

four eye-teeth, and twelve incisive . . . Age known by marks in mouth." (1854/1989, 6) (a description that does resemble eighteenth- and nineteenth-century taxonomies for the horse [Linne 1792; Donovan 1820], themselves influenced by the Aristotelian method of definition by genus and species)—are presented as harsh. Outside the restricting enclosure of the town, a different metaphor for knowledge is discernible:

> They walked on across the fields and down the shady lanes, sometimes getting over a fragment of a fence so rotten that it dropped at a touch of the foot, sometimes passing near a wreck of bricks and beams overgrown with grass, marking the site of deserted works. They followed paths and tracks, however slight. (Dickens 1854/1989, 353)

The value of an information system could lie in the ability it offers discriminatingly to follow "paths and tracks, however slight."

The etymology of intelligence offers further support for the significance of selection power. Traced to its Latin form (*inter-legere*: to choose from or between things), *intelligence* is strongly analogous to selection power, implying deliberate choice rather than domination by brute needs. Plutarch's account of the formation of the Roman military legion is better known:

> When the city was built, in the first place, Romulus divided all the multitude that were of age to bear arms into military companies, each company consisting of three thousand footmen and three hundred horsemen. Such a company was called a "legion," because the warlike were *selected* out of all. (1914, 123)

Thus, division or differentiation of individuals represents a further characteristic of man in the *polis* in its initial realization as the city-state. The current discussion conceives intelligence as a quality of human consciousness rather than inhering in the objects differentiated.

Comments on information systems by contemporary ordinary discourse are highly significant but difficult to produce as evidence. Evaluative criteria may be given by implication rather than fully and directly articulated. Yet a searcher who complains that it is difficult to control the number of records retrieved invokes a principle of discriminatory power. More explicitly, one spoken response to a presentation of the value of selection power was, "that's the basis [an enhanced capacity for informed choice] on which people use systems anyway" (Warner 2000, 79). Not influenced directly by the query transformation tradition, extradisciplinary written

comments can be regarded as embodying ordinary discourse concepts. For instance, a sociological study of communication in philosophy, which stresses the importance of direct oral communication between significant philosophers, criticizes literature discovery accomplished by "indexing and abstracting services (whether in printed media or electronically online), which overload the channels rather than focusing them" (Collins 1998, 45). Therefore, ordinary discourse concepts—although elusive—support the value of selection power.

The query transformation tradition of classic information-retrieval research also values discriminatory power, but for a distinguishable end that can be absorbed into selection power. The current argument values the discriminatory power of an index term for its potential to directly increase human selection and discrimination. Classic information-retrieval research, by contrast, has valued discriminatory power as a factor that assists automatic partitioning of documents or records (Van Rijsbergen 1979, 29, 136). The valuing of discriminatory power by classic information-retrieval research constitutes partly independent support for the value given to it here. The distinction in the ends for which it is valued reveals that this aspect of classic information-retrieval research can be absorbed into selection power as a special and incomplete case. The absorption of particular aspects of query transformation further indicates that the research tradition as a whole can be conceptually and operationally subsumed into selection power as a special case within a more encompassing theory. Therefore, this book will not pursue opposition with query transformation, which served as one dialectical point of departure for asserting the value of selection power (Wilson 1996c; Warner 2000).

Particularly since the mid-1990s, studies within information science have advocated selection, evaluation, and filtering as appropriate aims for information retrieval, rather than recall of all and only all relevant documents (Wilson 1996b; 1996c). The concept still needs to be fully operationalized (Griesbaum 2000). In relation to real-world considerations of human labor, the costs and value of that labor and labor delegated to technologies—the subsequent development of a labor theoretic approach to information retrieval—can be regarded as operationalization of selection power.

Summary

Analogous concepts in partly separate scholarly discourses support the value of selection power. Librarianship held a comparable concept in bibliographic control, indexing endorsed the discriminatory power of a term (a crucial factor in obtaining control), and cybernetics valued control and navigation. In addition, the value of selection power was supported by

- disputes within philosophy;
- a fictional critique of hierarchical classifications;
- the etymology of intelligence;
- and more specifically in relation to information retrieval systems, ordinary discourse comments on information retrieval.

The concept of selection power was elucidated and understood as the human faculty for discrimination; we endorse it here as the primary design and evaluation principle for information systems. As a special case of implementation, selection power absorbed query transformation, appropriate in certain circumstances and compelled by certain forms of technology (for instance by the batch-processing methods of the 1950s).

Thus, the first proposition for developing a labor theoretic approach to information retrieval asserts the value of selection power both as a property of human consciousness and as a primitive term, open to elucidation but not further decomposition.

Selection Power and Selection Labor

The relation between selection power and selection labor is not conceived either ahistorically or independently of technology. For instance, in primarily oral societies, forms of recitation not equivalent to verbatim repetition but still dependent on individual memory within a communal context were crucial to knowledge preservation by renewal (Goody and Watt 1968). The Icelandic law-speaker provides a transitional form that both inherits elements from orality and anticipates characteristics of written literacy. The law-speaker was required to recite the law and to answer queries on legal and parliamentary procedures by oral pronouncements based on his memory of the law (Njal 1280/1960, 306–308). From a modern perspective, the law-speaker could be regarded as an information

Box 2.1
An Information System Embodied in a Single Individual

Figure 2.3
Alþing in session. W. G. Collingwood.

Lawspeakers are characteristic of oral societies, and the Icelandic lawspeaker is of relatively late date and well-documented, with some concurrent and developing elements of written literacy. The lawspeaker would recite the law to the members of the annual assembly, or Alþing, as a linear spoken utterance, reciting one third of the laws at each meeting (Short 2008).

Selection power is evidenced in dialogic questioning and response, with evidence for the possibility of questioning separately from the recitation of the law. For instance, in *Njal's Saga*, the lawspeaker is consulted for confirmation of an aspect of the law.

> Flosi asked if this were the law, but Eyjolf replied that he did not know for certain and said that the Law-Speaker would have to settle that point. Thorkel Geitisson went on their behalf and told the Law-Speaker the situation, and asked if there were any legal basis for Mord's submission.
> "There are more great lawyers alive today than I thought," replied Skapti. "I can tell you that this is so precisely correct that not a single objection can be raised against it. But I had thought that I was the only person who knew this specialty of the law now that Njal is dead, for to the best of my knowledge he was the only other man who knew it." (Njal 1280/1960, 308).

Box 2.1
(Continued)

Selection labor can be discovered in the lawspeaker's mental work of memorization and recall, the communicative labor of recitation, and in the attention of the auditors. There is no direct analogue to products of description labor or metadata, further suggesting that these are historically specific developments of written literacy.

Technology is manifested in a natural object adapted to a human purpose, the rock face used as a sounding board for the voice of the lawspeaker:

> All those things which labour merely separates from their immediate connection with their environment are objects of labour spontaneously provided by nature, such as fish caught and separated from their natural element, namely water, timber felled in virgin forests, and ores extracted from their veins. (Marx 1867/1976, 284).

Therefore, selection power, selection labor, and technology are discernible in a primarily oral information system and have persisted across written literate and computational modes, indicating their centrality to information retrieval.

system embodied in an individual. With increased social complexity and the growth of both documents and indexes to documents, direct mental labor in memory and recall is transferred to sources outside the human body and mind—*exosomatic* resources. Variously conceived, the knowledgeable person may remain significant to information seeking and offer a readiness and focus of response difficult to obtain from more formalized information systems. In a development concordant with other features of secondary orality (Ong 1982), selection labor may be reemerging as a single category, concentrated in searcher labor.

Although acknowledged as a form of labor, even if not subjectively experienced as such, our primary concern will not focus directly on the mental labor of memory, recall, and response, but rather on the technological forms that now absorb the cognitive burden of memory and recall and also on the mental labor involved in their construction and searching. In premodern practice (that is, written and printed forms distinguished from computer-based, or modern systems), the organization of documentary materials required physical labor and, most significantly, the distinction between description and search labor was relatively clear. We will

carry this distinction forward as an analytical distinction, while recognizing the difficulty of substantive separation. For example, a searcher requesting a list of documents in chronological order effectively instantiates at the point of searching what once would have been a form of description labor, or process. The contrast is traceable back to the fixity of writing technologies compared with the possible fluidity of computation (Warner 2001, 33–46). The concern here is with the theoretical possibilities for and constraints on selection labor, conceived as incorporating both description and search labor.

Theoretical minima for selection labor can be derived from serial possibilities. If items are examined serially and without regress, selection labor increases with the number of objects examined in the collection. An unchanged principle for discrimination is assumed, rather than a conversational or dialogic alteration of the principle for discrimination, and this would be closely analogous to the batch processing practiced historically. An absence of meaningful organization of the discriminated objects and a conflation and simultaneous occurrence of description and search labor are also implied. If the choice between objects examined is reduced to a binary contrast between acceptance and nonacceptance, the unit of labor is somewhat analogous with the classic understanding of the bit.

Imposing organization upon a collection of objects separates description from search labor, with work invested in organization or description reducing labor and enhancing power in searching. Selection labor is then distributed between description and searching. If description and organization can partition objects into appropriate sets, Shannon's formula for the information of a source would indicate that the number of objects discriminated by search labor would rise more than the quantity of search labor:

> If there are N possibilities [for the choice of messages from a source], all equally likely, the amount of information is given by $\log_2 N$. . . . If it were possible to choose questions which always had the effect of subdividing into two equal groups, it would be possible to isolate, in twenty questions, one object from approximately 1,000,000 possibilities. (1968/1993, 214–215)

On this basis, the number of choices—broadly, units of labor—required to discriminate between c.1000 and c.1,000,000 such possibilities (which could correspond to objects, documents, or records for documents) would

double. The search could be conducted either deterministically or nondeterministically with human intervention at intervals, both with unaltered criteria during the process. The labor invested in description corresponds to a capital cost that is not incurred for each iteration of searching.

These slightly abstract considerations are helpful in establishing theoretical constraints for the labor associated with selection and also for enforcing the point that labor can be distributed between description and searching but cannot be eliminated. They also have some more practical resonances. The purposes of description may require semantic primitives that are difficult to isolate and may not exist, particularly for human or social discourse. An approach to theoretical limits can be made in other aspects of information theory, such as reducing redundancy in messages, but they are difficult to obtain fully (Shannon 1948/1993, 39; Verdú and McLaughlin 2000). Biological classifications might offer the closest analogies to reduction to atomic facts or to a perfectly organized source, but distinguishing species from variety can be problematic (Darwin 1859/1968, 104–108). Despite these reservations about the possibility of approaching theoretical limits on the effectiveness of descriptions, the

Box 2.2
Darwin on the Difficulty of Establishing Classifications

> Many years ago, when comparing, and seeing others compare, the birds from the separate islands of the Galapagos Archipelago, both one with another, and with those from the American mainland, I was much struck by how entirely vague and arbitrary is the distinction between species and varieties. On the islets of the little Madeira group there are many insects which are characterized as varieties in Mr. Wollaston's admirable work, but which it cannot be doubted would be ranked as distinct species by many entomologists. Even Ireland has a few animals, now generally recorded as varieties, but which have been ranked as species by some zoologists. . . . From these remarks it will be seen that I look at the term species, as one arbitrarily given for the sake of convenience to a set of individuals closely resembling each other, and that it does not essentially differ from the term variety, which is given to less distinct and more fluctuating forms. The term variety, again, in comparison with mere individual differences, is also applied arbitrarily, and for mere convenience sake.
>
> —Charles Darwin. *The Origin of Species by Means of Natural Selection or The Preservation of Favoured Races in the Struggle for Life*. 1859. (1859/1968, 104–108).

potential that the increase in the number of objects discriminated will be much greater than the increase in the search labor required for discrimination might provide an underlying explanation for the possibilities of scaling offered by practical devices such as thesauri in working information systems.

Therefore, the second proposition for developing a labor theoretic approach to information retrieval is that selection labor produces selection power, both for an oral speaker and with certain forms of exosomatic technologies, which tend to be adopted under historical circumstances of increased social complexity. Selection labor's production of selection power and the decomposition of selection labor into description and search labor is very clearly exemplified within written literacy, but continues in modified form with modern information technologies.

Conclusion

This chapter addressed some fundamental issues and developed some propositions. Information retrieval research was reviewed and selection power was endorsed as the primary value or aim for information retrieval. Selection power was received as a primitive term whose content was elucidated, and value supported, from analogous concepts in partly independent discourses and also from its embodiment in information retrieval practice. Selection power was produced by selection labor. An understanding of information retrieval has begun to absorb the concept of informational or mental labor. There is a congruence of values with processes for information retrieval, particularly through the common idea of selection.

For the further development of a labor theoretic approach to information retrieval, the concept of selection labor requires further differentiation to develop and exemplify the distinction between description and search labor and explore the possibility of transferring direct human labor to technology. Semantic mental labor will be fully distinguished from syntactic mental labor, differentiating mental labor from the processes that can be abstracted from it and transferred to technology, and from the products that can result from mental labor. These topics will be the concern of the next chapter.

3

Description and Search Labor

Introduction

The previous chapter characterized selection labor as a form of mental labor and established theoretical minima for a given collection of objects. It also noted that the premodern technologies of writing and printing on paper separated selection labor into description and search labor. Similarly, the discussion acknowledged the reconvergence of description and search activities with computer-based, or modern, technologies, together with the possibility of sustaining analytical distinctions between them. The activities of description and searching still require more fully empirical characterization as components of selection labor.

This chapter will develop and elucidate the concepts of description and search labor. As mental labor, description and search labor participate in the conditions for other forms of mental labor. Therefore, the chapter will introduce distinctions between types of mental labor and their different possibilities for transfer to technology. The perspective established on mental labor and technology is then used to understand description and search labor and their relation to selection labor.

Concepts of Mental Labor

Labor and Mental Labor

For both *Genesis* and Marx, labor—productive work in nature—is a fundamental condition of human life, imposed by the necessity for survival.

> The labour process . . . is purposeful activity aimed at the production of use-values. It is an appropriation of what exists in nature for the requirements of man.

> It is the universal condition for the metabolic interaction (*Stoffwechsel*) between man and nature, the everlasting nature imposed condition of human existence, and it is therefore independent of every form of that existence, or rather it is common to all forms of society in which human beings live. (Marx 1867/1976, 290)

Mental labor is usually conceived as an adjunct to enhancing control over the physical environment, but it also can be considered as an activity in itself, with possibilities for mechanization explored (Minsky 1967, 2).

Agrarian, industrial, and information technologies can be regarded as human constructions, the product of human labor on natural resources and preexisting human-made products, rather than naturally or objectively given. In Marx's work, the understanding of technology as a human construction receives its fullest expression in a classic passage from the *Grundrisse*, whose themes implicitly inform the treatment of technology in *Capital* (Marx 1867/1976).

> Nature builds no machines, no locomotives, railways, electric telegraphs, self-acting mules etc. These are products of human industry; natural material transformed into organs of the human will over nature, or of human participation in nature. They are *organs of the human brain, created by the human hand*; the power of knowledge, objectified. The development of fixed capital indicates to what degree general social knowledge has become a *direct force of production*. (Marx 1858/1973, 706)

Although Marx mentions automatic devices involving control mechanisms ("self-acting mules" [automatic spinning machines]), the passage focuses on industrial technologies common during his historical period. Regarding information technology as a human construction and concerned with the transformation of signals rather than natural resources (Warner 2004, 5–35), indicates the possibility of a similar transfer of human mental labor to technology. For both industrial and information technology, transferring direct human labor to technology can speed processes and enable previously impossible activities.

The possibility of transferring aspects of direct human labor to technology enables a dynamic between human labor and its technological products, in which forms of direct human work are progressively transferred to technological processes. The dynamic is compelled by the historical search for greater control over the environment and accelerated by the innovatory dynamic of capitalism, including the impulse from reduced immediate costs of labor transferred to technology. From the perspective

indicated here, the dynamic can apply to mental labor as an independent activity, not simply an adjunct to control the physical environment. The social division of labor (for instance, between master and slave, or intellectual and clerical labor) can anticipate the division of labor between human labor and machine process.

We can distinguish different types of labor, adopting Marx's distinction of universal and communal labor and applying it to informational labor, processes, and products.

> We must distinguish here, incidentally, between universal labour and communal labour. They both play their part in the production process, and merge into one another, but they are each different as well. Universal labour is all scientific work, all discovery and invention. It is brought about partly by the cooperation of men now living, but partly also by building on earlier work. Communal labour, however, simply involves the direct cooperation of individuals. (1894/1981, 199)

As they merge into each other, universal and communal labor are not encountered in pure form and can be embodied in the products of labor, but their distribution can vary significantly. Communal labor should also be understood to include individual labor. In this context, we can regard

- the machine aspects of information technology primarily as products of universal labor, although constructed, activated, and renewed by communal labor;
- the design and writing of programs as communal labor developed from universal labor and its products, such as understandings of the algorithmic process and programming languages; and
- human description of information objects as primarily involving communal labor, guided by universal labor and embodied in codes for description.

Communal labor has immediate costs—possibly matched by a wage—in required human energy; in contrast, the costs of universal labor have been absorbed over time. As formulated by Marx, the distinction between communal labor and universal labor, involves mental labor—"Universal labour is all scientific work, all discovery and invention" (Marx 1894/1981, 199)—and also can be applied directly to the modern manifestations of mental labor and its products.

Labor, process, and product can be distinguished explicitly—process and product separated from an originally undifferentiated labor. For

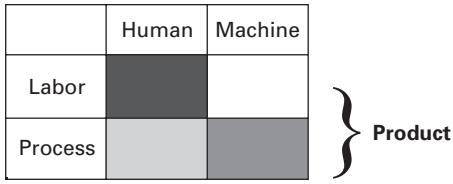


Figure 3.1
Labor, process, and product.

instance, in unrecorded oral speech, labor and process are not necessarily distinguished and the product disappears with its creation. Once the categories are separated, labor can be understood as direct human work and process as activity abstracted from labor, with the possibility of transfer to technology. This partly nominal distinction of labor from process, which arose originally from unease at applying the term *labor* to inorganic activities, also has substantive implications: compared to direct human labor, greater rigidity and exactness in the process may arise partly from preliminary formalization. Products—in this context, semiotic products such as catalog records—result from direct human labor and technological processes.

Figure 3.1 diagrams the distinction between labor, process, and product. Labor is exclusively human—indicated by the darkly shaded intersection of *Human* and *Labor*—while the white box indicates the absence of intersection between labor and machine. In contrast, process can be human or machine. Human labor can be formalized as a process (intermediate grey section) and fully formalized processes can be transferred to technology (darker grey section). The making of products involves both labor and machine processes that may congeal in the product, requiring deliberate investigation to discover their roles in the making of the product.

An object for description can be read in as a precondition for bibliographic labor, processes, and products (to the left of the diagram). The object for description could be a natural object or, more characteristically, a product of human mental and productive labor, such as a written document. Under premodernity, labor—embodied in objects for description—was usually received as a given, both in theory and in bibliographic description practice. Under modernity, full text searching can reactivate

the labor embodied in objects and transform it into a resource for exploitation.

We can now introduce distinctions specific to mental labor that are not captured by historical concern with physical and productive labor.

Distinctions within Mental Labor

The social division of intellectual labor is more familiar than separation between the aspects of mental tasks, first as a division between the priesthood and those engaged in more directly productive labor (Childe 1956), and developing with the differentiation of disciplines and discourse communities (Goody and Watt 1968).

The nineteenth century witnessed deliberate division of the aspects of mental tasks between people working either within organizations or on extensive projects, in an addition to the social division of mental labor. A theoretical understanding of the possibilities surrounding the division and mechanization of mental labor also began to emerge. At a level that mediates between practice and theory, the production of logarithm tables was determined amenable to manufacture by subdividing the tasks involved and dividing the human labor those tasks required. Inspired by Adam Smith's work on the division of labor, Prony conceived:

> l'expédient de mettre ses logarithmes en *manufacture* comme les épingles [the expedient of putting his logarithms into *manufacture* as if they were pins]. (Babbage 1835/1963, 193)

The actual process of producing the tables revealed an interesting and potentially disturbing feature: the relative accuracy of clerical and intellectual human labor.

> Persons [nine tenths with no knowledge of arithmetic beyond addition and subtraction] were usually found more correct in their calculations, than those who possessed a more extensive knowledge of the subject. (Babbage 1835/1963, 195).

Therefore, considerations of meaning for human computers can distract attention from simple computational operations. At a more deliberately theoretical level, Babbage, similarly influenced by and adapting Adam Smith on the division of labor in manufacturing, noted the possibility for the division of mental labor.

> The division of labour is no less applicable to mental productions than to those in which material bodies are concerned. (Babbage 1835/1963, 379)

Mental labor has a material aspect, particularly regarding the use of exosomatic technologies, and it is this material aspect that gives the possibility of mechanization.

Mechanical mental labor contrasts with mental labor belonging to "the domain of the understanding, requiring the intervention of reasoning" (Babbage 1989, 246) and closely involving considerations of meaning. For Babbage, one impulse to the design and construction of computing machines involved transferring the mechanical part of the mathematician's labor to automatic machinery (1989, 246–247). A similar impulse can be discerned in the diffusion by adoption of modern information technologies, a process also influenced by a desire to avoid the costs of direct human labor or, in Marx's terms, of communal labor.

The familiar—and deeply embedded—distinction of semantics from syntax can be used to differentiate semantic from syntactic mental labor. Semantic labor is concerned with transformations motivated by the meaning or significance of symbols, while syntactic labor is determined only by the form of symbols, operating on them in their aspect as signals. Semantic labor requires direct human involvement; by contrast, originally human syntactic labor can be transferred to information technology, where it becomes a machine process. Direct human labor has high costs; under modern conditions, mental labor transferred to technology likely will decrease costs. The distinction of semantic labor from syntactic labor has an analogue in ordinary discourse, and the transfer of syntactic mental labor to technology occurs in everyday practice, suggesting the robustness and wide applicability of the distinction, despite its only recent theoretical articulation (Warner 2005a).

The following literary example, occurring chronologically close to and within the same broad cultural context as Boole's formalization of logic (1854), conceived as the laws of thought, conveys the distinction between semantic and syntactic mental labor:

> "My other piece of advice, Copperfield," said Mr. Micawber, "you know. Annual income twenty pounds, annual expenditure nineteen nineteen and six, result happiness. Annual income twenty pounds, annual expenditure twenty pounds ought and six, result misery. The blossom is blighted, the leaf is withered, the God of day goes down upon the dreary scene, and—and in short you are for ever floored. As I am!" (Dickens 1850/1981, 175)

In this passage, the substitution of a semantically congruent but syntactically dissonant result parodies the syntactic process of calculation—a curious and revealing inversion of modern experience with computer operations on written text for information retrieval and spell checking. In contrast to the syntactically generated result it implicitly replaces, the semantic result does not generalize to other similar syntactic procedures. In nineteenth century practice, the exosomatic technologies of writing would have assisted human mental labor, and both semantic and syntactic labor would have involved continuous human intervention. Following the mechanization of mental labor during the late twentieth century, syntactic labor could be transferred to information technology—operating deterministically between intervals of human intervention and opening up and revealing a distinction between semantic and syntactic mental labor. Excepting processes already formalized or known to be formalizable, attempts to transfer human mental labor to information technology usually reveal the complexity and intractability of semantic labor.

The endorsement of strongly analogous distinctions—implicit in a highly significant copyright judgment by the U.S. Supreme Court involving modern information technologies—supports the distinction between semantic and syntactic labor and also between labor and process. In *Feist v. Rural*, the court denied copyright protection to telephone white pages:

> Nor can Rural claim originality in its coordination and arrangement of facts. The white pages do nothing more than list Rural's subscribers in alphabetical order. This arrangement may, technically speaking, owe its origin to Rural; no one disputes that Rural undertook the task of alphabetizing the names itself. But there is nothing remotely creative about arranging names alphabetically in a white pages directory. It is an age-old practice, firmly rooted in tradition and so commonplace that it has come to be expected as a matter of course. . . . It is not only unoriginal, it is practically inevitable. (Feist 1991, 363)

Age-old practice corresponds to syntactic labor and the alphabetization of names corresponds to a syntactic process, delegated to the forms of information technology current in the 1980s. Slightly obscured within the Court's judgment—and not fully brought into contrast with its explicit reversal of the labor theory of copyright—is a reference to "writings which are to be protected . . . [as] *the fruits of intellectual labor*, embodied in the form of books, prints, engravings, and the like" (1991, 346). The intellectual labor embodied in the protected writings is closely analogous

to semantic labor. While the judgment has independent interest, its immediate value in this context is as a wider public analogue to the distinctions of syntactic and semantic mental labor and of human labor from machine process, thus supporting their validity.

We can now summarize the understanding of the conditions for mental labor. Human labor, including mental labor, can be transferred to technology, undergoing transformation into a machine process. Human mental labor can be semantic or syntactic in character, directly motivated by considerations of meaning or reduced to pattern-governed transformations. Only syntactic, not semantic, labor can be transferred to information technology. As forms of mental labor, description labor and search labor participate in the conditions for mental labor.

Description and Search Labor

The historical separation (and current reconvergence) of description labor and search labor from selection labor supported the proposition that selection labor can be distributed significantly between description labor and search labor, but its overall sum could not diminish below certain theoretically established limits. The understanding of description labor and search labor as forms of mental labor supports the further proposition that the syntactic aspects of description and searching, but not their semantic components, can be transferred to technology.

Description Labor

Thus, description labor is understood as one component of selection labor. As mental labor, description labor can have a material and exosomatic aspect and semantic and syntactic components. We will consider description labor in a more directly empirical and contemporary fashion than selection power and selection labor, but the historical emergence of description labor in information systems with the emergence of written literacy is acknowledged here. We also acknowledge equally the possibility that description labor can be reduced, transferred to technology as process, and absorbed by selection labor (with selection labor becoming a more substantive category). The labor embodied in documents described is accepted largely as a given and not fully explored, but the contrast

between the technology of writing on paper, demanding separate description, and computation that enables automatically generated syntactic descriptions is fully incorporated into the schema developed. Human description labor includes interpretation as well as apparently more simple forms of description, classically understood as embodied in the activities of both classification and cataloging.

Description labor is understood ostensively and empirically as the work involved in processes such as cataloging, classification, indexing, and database description. More analytically, although still consistently, description labor is conceived as the work involved in transforming objects (documents, images, or people) into searchable descriptions that will assist subsequent retrieval.[1] Two aspects of the transformation of objects for description can be distinguished: the description of objects and the assembling of these descriptions into searchable lists or indexes. These aspects may merge into each other in practice, but they can be separated analytically. Description labor aims implicitly to increase selection power and can have the further effect of reducing labor expended in searching.

The object for description (see figure 3.2) is both physical document and ideational text—information as both object and potential knowledge (Buckland 1991; Blair 2002). As text, the object for description is the product of an extended period of human mental labor, with periods of high intensity: "writing a book is a horrible, exhausting struggle" (Orwell 1946/1970, 29). The purchase price of the document may not fully reflect the duration, intensity, and real costs of that labor.

The separation of syntax from semantics with the advent of written language (Warner 2005a, 557–563) may have continuing implications for the relative effectiveness of applying syntactic, or pattern-transforming, procedures to verbal and nonverbal graphic signs. Nonverbal signs can be subjected to syntactic or pattern-based transformations, with those transformations realized as labor or as machine process, but it has been difficult to endow transformations with semantic significance, either for similarity and difference between signs or for establishing orders for display. Technically, it would not be difficult to transform a digital photograph into a searchable description (for instance, Google's advanced image search function enables searching by a range of colors and file types [Google 2007b]) and to establish measures of similarity between images.

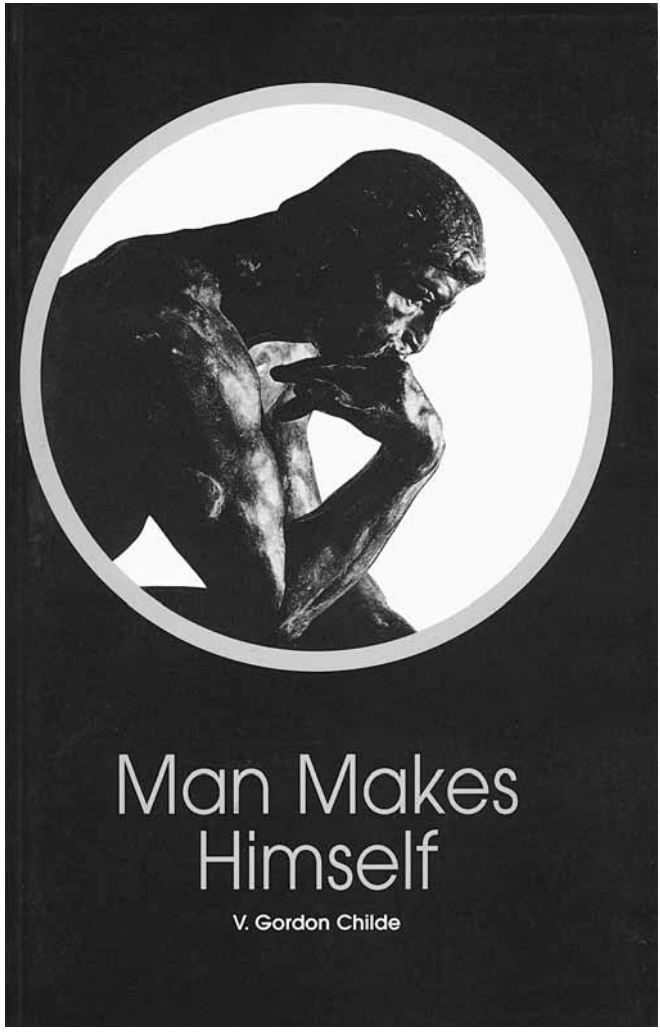


Figure 3.2
Object for description.

However, excepting sets of images produced under highly controlled conditions and thereby effectively selected for their formal similarity to one another, creating meaningful searchable descriptions or measures of similarity has been difficult. Images, including photographs, received as iconic signs with a motivated resemblance from signified to signifier could lend themselves to iconic representation for display and retrieval, possibly reducing image size to enable scanning. Google uses indexing and retrieval based primarily on verbal descriptions of images, not on matching between the graphic images; iconic modes of representation are used to display retrieved results (Google 2007b). In contrast, written verbal signs that embody the distinction between semantics and syntax are more amenable to syntactic operations, which can partially incorporate semantic significance.

A bibliographic record (see figure 3.3) represents one manifestation of the product of description labor and processes, potentially open to further transformations. Enhanced selection power, or search control, is made possible by reducing the diversity of language in related objects for description to canonical or standard forms. In many instances, categorizing objects implicitly or explicitly involves the relationship between individual (or species) and genus, gathering together items related in meaning. Collecting items related in meaning but diverse in pattern often involves human semantic labor. Under premodern conditions, enhanced control generally involves abstraction and loss of richness. Characteristically, the description is briefer than the object described and also is the product of less extensive and less intense labor, although the costs of that labor may be recovered more directly.

Examples of significant information systems can reveal the reduction of direct human labor and the increased fullness of descriptions. In late-nineteenth- and early-twentieth-century practice, *Palmer's Index to the Times Newspaper* was created by direct human labor and assisted by current technologies, particularly writing and printing. For instance, "Mad Dogs," a *Times* subheading that introduced a short paragraph beginning, "The inhabitants of Sheffield . . . ," would be transformed into the index entry "Mad dogs in Sheffield," and filed under "M" in the list of entries for that volume of the index (Morrison 1986, 189–192; *Times* 1885, 6; Palmer 1885, 60). The index, as a product and as a whole, can be read to reveal

Man makes himself /

V Gordon **Childe**

2003
English Book 244 p. : ill. ; 22 cm.
Nottingham : Spokesman, ; ISBN: 0851246494 9780851246499

	GET THIS ITEM
Availability:	**Check the catalogs in your library.** ∞ Libraries worldwide that own item: 15 ∞ Connect to the catalog at your library
External Resources:	∞ QConnect ∞ Cite This Item
	FIND RELATED
More Like This:	Search for versions with same title and author | Advanced options ...
Find Items About:	Childe, V. Gordon (max: 29)
Title:	**Man makes himself /**
Author(s):	Childe, V. Gordon 1892-1957. (Vere Gordon),
Publication:	Nottingham : Spokesman,
Year:	2003
Description:	244 p. : ill. ; 22 cm.
Language:	English
Series:	New thinker's library ; no. 2;
Standard No:	**ISBN:** 0851246494; 9780851246499
	SUBJECT(S)
Descriptor:	Civilization, Ancient. Archaeology. Prehistoric peoples. Progress.
Note(s):	Reprint of the 4th ed./ Includes bibliographical references (p. 239-240) and index.
Class Descriptors:	**LC:** CB311
Responsibility:	by V. Gordon Childe ; with a new foreword by Mark Edmonds.
Vendor Info:	YBP Library Services Baker & Taylor (YANK BKTY) 42.50 **Status:** active
Document Type:	Book
Entry:	20040730
Update:	20071103
Accession No:	**OCLC:** 56060701
Database:	WorldCat

Figure 3.3
Product of description labor and processes. Source: OCLC. WorldCat 2008.

that the description process was strongly syntactic, although conducted by human labor rather than by machine, with some semantic intervention. The storage constraints of the technologies compel the production of descriptions more concise than the objects described. Technologies also have to be used in a primarily nondeterministic mode with continuous human intervention, particularly for construction rather than printing of indexes. Other information systems of the time (most obviously the construction of the *British Museum Catalog*) involved substantial human semantic intervention in the making and listing of descriptions—of material by and about significant authors—although the subject approach was not fully endorsed (Roberts 1977, 14).

These contrasting descriptive practices have modern descendants that can be embodied within the same information system. Google's advanced search descends from the practice of indexing newspapers by syntactic transformations on object-language. Further semantic work is not applied to the objects described, although they may contain deliberate descriptive elements (for instance, in the application of metadata). Descriptions not available for direct inspection are automatically generated from the verbal objects themselves and further generate indexes to those descriptions. Direct human clerical labor that was involved in nineteenth-century practice is now delegated to modern information technologies that operate syntactically and deterministically. The need for briefer descriptions imposed by storage constraints has also been removed, but the searcher's labor remains intense.

WorldCat embodies both practices. Guided by codes developed from their historical antecedents, descriptions are created by human semantic work, with the aim of creating more systematic descriptions than those given by transcription of the verbal objects. These systematic descriptions aim to increase the searcher's selection power and reduce labor. Elements of syntactic labor can also be found, transformed into a machine process. For instance, *WorldCat* automatically transforms humanly created descriptions into searchable indexes. Descriptions themselves increasingly incorporate information derived by syntactic processes from the verbal objects taken for description.

Some common trends over time can be discerned in these processes of description. There is a decrease in the direct human labor involved in the

Box 3.1
Nineteenth-Century Identity Description Practices

Description labor, processes, and products can be realized and discerned in activities other than cataloging and indexing. For instance, descriptions of people for the purposes of controlling travel and citizenship have intensified since the developments in global communications (both physical transport and message transmission) in the mid- to late-19th century. Identity descriptions of people and descriptions of documents may share a similar but more explicitly conceived aim to identify and differentiate objects for descriptions (for instance, individuals as citizens). Jules Verne gives a gently ironic account of the process of acquiring a passport in Paris in 1859:

> He then went to the prefecture of the Seine—for Jacques 'the lord mayor's parlour'—where he boldly requested a passport for the British Isles. His description was taken down by an old, myopic clerk whom the progress of civilization would one day replace by an officially designated photographer. (1992, 6–7).

Photographs are a more effectively translingual but not culturally invariant means for recognizing identity than verbal descriptions.

Prior to the use of photographic images for establishing identity, or in the empirical absence of an identifying image from a particular situation, other behavioral characteristics were used to identify individuals or to differentiate them from other individuals with whom they were or could be confused. In Thomas Hardy's "The Three Strangers," an escaped prisoner, who had been due to be hanged, is identified by his "musical bass voice that if you heard it once you'd never mistake as long as you lived." (1888/1976, 34). In O. Henry's "The Theory and the Hound," a wife murderer—whose verbal description could encompass his companion even though "[t]hey bore little resemblance one to the other in detail,"—is differentiated by his highly angry response to the investigating sheriff's deliberate cruelty to an animal: "I never yet saw a man that was over fond of horses and dogs but what was cruel to women" (1910/1993, 401, 406). Both narratives reveal an ordinary discourse understanding of the crucial value ascribed to an identifying or discriminatory variable.

In modern practice, description processes in many situations are delegated largely to technology for the production of images, requiring human semantic labor for recognition or comparison of image and person. Consider, for instance, the making of images and their interpretation by airport security systems.

description of objects, and, more intensely, in the compilation of indexes from these descriptions. In particular, syntactic components of description are being transferred from human labor to machine process. Descriptions are becoming more full and exact, enabled but not compelled by the

reduction in technologically imposed storage constraints. Fullness and exactness can be distinguished analytically and substantively from discriminatory power and informativeness; they may not enhance selection power, but they may aid specificity in searching.

The social division of mental labor between clerical and intellectual roles has been transformed progressively into the division of labor between human work and machine processes, with syntactic labor delegated to technology. While contrasts between late-nineteenth-century and current practice arise from the development and adoption of information technologies—the progressive transformation of communal labor into the products of primarily universal labor—continuities follow from the predominantly symbolic nature of the signs described (Warner 2005b).

A particular feature that combines continuity and change involves transformation from social division of labor into division of labor between human and machine, with syntactic labor transferred to machine process. This pattern might have implications for the nature of expertise in searching information systems. At one extreme, understanding the forces that determine a system's construction may not correlate with effective searching of systems.[2] The changed relation between language of discourse and language of representation between premodern and modern information systems is acutely relevant: a semantically assigned language of representation need not be the entry vocabulary, although later recourse could be made to it for its generic scope. *Prima facie*, expertise in the language of discourse likely will be stronger among domain than information specialists.

A few other writers have considered the costs of human description labor involved in creating records for catalogs, although not fully within the conceptual framework developed here, and these considerations can provide empirical data to inform the argument. The costs associated with creating a catalog that meets *WorldCat* standards have been estimated at about US$40 (Hayes 2000, 76). This amount represents the cost of human semantic description labor, and the use of *WorldCat* records by participating libraries represents distribution of the products of labor. Online Computer Library Center (OCLC) accounts show no value, either for *WorldCat* or for OCLC's direct costs for creating the records (Hayes 2000, 76). Failure to recognize the cost of direct and accumulated intellectual labor embodied in catalog records and catalogs represents a "serious

mis-measurement of phenomena of fundamental economic importance" (Hayes 2000, 73).

In summary, description labor can be semantic or syntactic in character. Semantic description labor is exclusively and directly human, in accordance with the understanding of human mental labor developed in this chapter and the possibilities of its transfer to technology. In contrast, syntactic description labor can be delegated to technology, which transforms it from organic labor into machine process.

Search Labor

Search labor was contradistinguished from description labor; both description and search labor are regarded as components of selection labor. Search labor is also a form of mental labor, containing material, semantic, and syntactic aspects. Within the schema developed, selection labor distributes significantly between description and search labor and its syntactic aspects are transferred to technology. Search labor might have emerged predominantly as search expertise in premodern systems, reflecting the substantial investment of human semantic labor in description processes. A signal and revealing exception to the emergence of labor primarily as expertise is represented by the arduous work and accumulating expertise required for searching for the sources necessary to construct a subject bibliography, attempting exhaustivity and encountering bibliographic scatter (Greg 1959; Bradford 1948/1971).[3] For analytical clarity and with substantive justification, search labor is the reverse, obverse, or mirror image of description labor.

Search labor is reflected in its material aspect in the psychomotor skills required to operate a keyboard, mouse, or other interface devices. The progressive naturalization of information technologies and improved interface resulting from system design have diminished the apparent complexity of this aspect of searching. Both perceived and inherent complexities have diminished, although human desires may evolve alongside technology. The recent relative stability of interface technologies may indicate that the late-twentieth-century revolution in the mechanization of mental labor resulted in a teleological state, at least temporarily.

Syntactic processes can be distinguished from syntactic aspects of searching. Derived from processes previously conducted directly by

humans, syntactic processes now have been transferred to machines. In contrast, syntactic aspects of searching involve human understanding of syntactic processes. Syntactic processes—activities such as ordering retrieved records or eliminating duplicate records—increasingly are transferred to technology. With premodern technologies, ordering records by date or by author would have been accomplished at the point of retrieval by description labor or direct human syntactic labor. Modern search technologies can invoke different orderings, transferring human labor to machine processes.

Compared with syntactic processes, the syntactic aspects of searching correspond to human understanding of syntax. Specific components for understanding include Boolean logical combinations, their computational derivatives, their realization in system commands, and the likely effects of specific combinations. Many consider Boolean logic difficult to grasp, although its difficulty may have been exaggerated.[4] The adoption of computational technologies has resulted in some public understanding of Boolean logic. Boolean operators are characteristically realized in contrasting system commands, although the underlying commonality should be recalled. Early studies indicated that the number of system commands used had little effect on system performance (Barraclough 1977). However, differences in commands made it difficult for searchers to adapt from one system to another. Similar to the material aspects of searching, syntactic aspects have improved through simpler system design, and also through coevolving consciousness.

Semantic components center on translating a topic into a searchable query, the most complex aspect of retrieval (Roberts 1977, 15; 1989). The process may remain difficult, but its nature—the social distribution of expertise and the distribution of labor between description and searching—may be changing due to alterations in description processes. Semantic understanding and expertise with premodern technologies involved understanding the type and particular characteristics of description labor applied to those documents, including the language of subject representation. With modern systems, understanding the documents' language of discourse and the effects of generating automatic representations of those documents may become more significant, likely involving a social redistribution of expertise toward those fully familiar with the language of discourse.

A further level of semantic understanding involves knowledge of the overall system of information system production and of the role or position of particular information systems used within that system. The term *system* refers to an interacting set of components rather than a deliberately coordinated or planned system (Roberts 1977). The diffusion of Internet search engines is changing both the overall system and the nature of expertise.

A significant contrast between description and search labor lies between the presence of the object for description and the customary initial absence of the object in searching. Presence informs description and absence combined with desire to restore presence motivates searching. Objects or documents recalled in search can be received as exemplar documents and used (possibly iteratively) to modify the search regarding choice of terms from the language of representation, where it exists and can be exploited, and the language of discourse. The presence and absence of the object creates a particular form of abstraction or disconnection for machine description processes and human searching of the products of those processes. In machine description, words are characteristically torn or abstracted from their context, understood specifically as the line of writing. They may be considered apart from their lines of writing by a human searcher, but machine processes restore their various lines of writing during retrieval. They are then made more fully present and subjected again to human semantic interpretation, informed by their line of writing and their broader context (see chapters 5, 6, 7, and 8 for fuller discussion of these patterns). Theoretically, it is possible to reconstruct documents from full-text descriptions, at least as linear sequences of words. Thus, the contrast between the presence and absence of the object described initially distinguishes description from searching: presence informs description and absence motivates searching. The contrast tends to produce a mirror image of description in searching.

Therefore, search labor parallels description labor, and search labor is either semantic or syntactic in character. Semantic search labor is inescapably and directly human, while syntactic search labor can be transferred to technology, where it becomes a machine process. Obtained without a Procrustean fitting of data to theory and without disguising their contrasts, the parallelism between search labor and description labor rein-

forces the analogies between them. It also strengthens the possibility of interchanging human labor and machine process between description and searching within the encompassing activity of selection labor.

Dynamic for Transfer to Technology

The dynamic that compels the transfer of syntactic labor to technology is connected with the costs of direct human labor, and was acutely noted in a proleptic remark by Norbert Wiener.

> the automatic machine . . . is the precise economic equivalent of slave labor. Any labor which competes with slave labor must accept the economic conditions of slave labor. (1954, 162)

As semantic description labor decreases, semantic labor may transfer from description to searching and search labor may increase. Thus, the process of searching may remain permanently intractable at the semantic level.

Conclusion

This chapter elucidated the concepts of description and search labor and also provided parallel distinctions for the work involved in description and searching. It established the possibility of transferring human mental labor to technology. The discussion characterized semantic labor as irreducibly human and syntactic mental labor as transferable to technology as machine process. Description labor involves cataloging, classification, and database description, with separate semantic and syntactic aspects. Search labor strongly parallels description labor. Selection labor, or the total labor involved in information retrieval, is amenable to significant distribution between description and searching, and its syntactic aspects can be transferred to technology.

A further stage in the development of a labor theoretic approach to information retrieval involves reviewing and synthesizing all the elements of our argument—from selection power to selection, description, and search labor—and fully introducing the costs of direct human labor as the dynamic that compels the transfer of human syntactic labor to technology. The parsimony and power of the overall argument and its correspondence to real-world decision practice will also be revealed.

4

A Labor Theoretic Approach

Introduction

Although the previous chapters established the essential components of a labor theoretic approach to information retrieval, a synthesis is still needed. The synthesis should reveal the economy and power of the approach by emphasizing the coherence and mutual relation of elements such as selection power and selection labor. Emerging from the costs of direct human mental labor and a common desire to avoid those costs, the practices of major information systems offer an empirical illustration of the dynamic that transfers human mental labor to machine processes. Decision practices of major information systems embody and conform to the strongly determining forces already identified. This chapter will consider the value of the labor theoretic approach and suggest the continuing intractability of information retrieval.

A full synthesis of the labor theoretic approach must recapitulate its discursive development and lay bare its logical structure and progression. The limited number of concepts and activities distinguished reveals the economy of the approach, but its power lies in its ability to comprehend significant empirical developments in major information systems and selectively absorb preexisting theories.

The argument presented in the previous two chapters was a logical, but not reductionist, approach. Concepts such as selection power and selection labor were abstracted from everyday practice—from Marx's "real life process" (1858/1973, 706)—rather than imposed theoretically. Thus far, we have preserved and incorporated the empirical richness of information retrieval into a discursively expressed cohesive structure. This chapter

will introduce formal (symbolic) logic to expose the underlying rigor of the argument. Connectives from formal logic are fully sufficient to capture relations between the "atomic fact[s]" distinguished (Wittgenstein 1922/1981, § 2.062); atomic facts are understood to include both concepts and activities (for instance, selection power as a concept and selection labor as an activity). By embedding formal logical exposition in a summary discursive recapitulation and complementing it with a diagram, we will sustain intelligibility for readers familiar with ordinary discourse but not fully conversant with formalism. Thus, complementing the discursive development with a more formal presentation should clarify and reveal the rigor of the argument and also maintain strong correspondence to reality.

Synthesis

The first concept we treated was selection power, received as an atomic fact or primitive term that is open to elucidation but not definition in the sense of decomposition into more primitive terms. Selection power was understood as human faculty for discrimination, which could be augmented by technologies but still remain a quality of human consciousness. Therefore, our first proposition asserted selection power's value for information retrieval.

Labor was regarded as comparably fundamental to selection or intelligence and an inescapable condition of human existence. Labor was further understood to include mental as well as physical labor, both as an adjunct to physical or productive labor and as a separate activity. A specific concern involved selection labor as a form of mental labor. Thus, selection labor could also be asserted as a primitive activity, although inescapably imposed by the need for selection power and not implicitly or explicitly sought after as a good.

We have developed a central proposition connecting the primitive concept and activity—that selection labor produces selection power under a range of historical conditions: emerging when orality transformed into literacy; fully realized with literacy, or premodern, information technologies; and continuing, although transformed, with modern technologies. This proposition has empirical correlates in historical and current description processes for documents, objects, and people. Under premodern con-

ditions, selection labor decomposed clearly and strongly into description and search labor.

The production of selection power by selection labor is analogous to material implication in formal logic, with selection power implying selection labor.

Selection power → selection labor.

Making an explicit analogy that links the production of selection power by selection labor with the relation of material implication in formal logic adds value to the discursive exposition. The relation of material implication remains valid for selection labor without selection power[1] and then caters to the case, met in practice, that selection labor does not produce selection power when unhelpful or inappropriate selection labor has been applied. In material implication, the first place given to selection power also suggests that it is possible to hide or disguise selection labor from view behind selection power (for instance, congealed in the products of selection labor). Disguising selection labor also corresponds to its theoretical neglect, despite costs and significance (Hayes 2000), and may be partly responsible for that neglect. In the practical use of a particular system, a searcher might have to determine the nature of selection labor or process applied from the selection power afforded; for instance, variants and unreconciled forms of a single author's name could indicate the absence of effective human description labor or machine processes in reconciling variants.

Some further propositions illuminated the relation of selection labor to description and search labor and the possibilities of their transfer to technology. Concepts isolated and distinctions made corresponded to real-world practice, suggesting their robustness. Labor was understood to include mental labor.

Labor includes mental labor.

Mental labor shared some characteristics with physical and productive labor. Direct human mental labor could be transferred to technology, and technology itself was regarded as a human construction.

Human mental labor *can be transferred to* information technology.

In the intervals between direct human intervention, fully transferred human mental labor becomes a machine process.

Some categories developed primarily for physical and productive labor explicitly recognized elements of mental labor. According to Marx, universal labor—"all scientific work, all discovery and invention"—differentiates from communal labor: universal labor is "brought about partly by the cooperation of men now living, but partly also by building on earlier work," while communal labor "simply involves the direct cooperation of individuals" (1894/1981, 199). Labor might originate as communal labor, but it is transformed into universal labor as labor-created knowledge diffuses.

Labor *separates into* communal labor and universal labor.

Communal labor includes individual labor. In both hardware and software, information technology results primarily from universal labor. Human descriptions of information objects (records) result predominantly from communal labor.

Distinctions of labor, process, and product, understood close to their ordinary discourse senses, could also be applied to mental labor and its products. In oral speech, particularly under orality that preceded the development of written language, labor includes the entire activity of communication and process and product are not fully separated. In written language, the product is separate from labor and the possibility emerges of formalizing the processes involved in making that product. Under modernity, processes can be conducted automatically. Therefore, originally undifferentiated labor separates into labor, process, and product, with associated changes in the meanings of the constituent terms, or:

Labor *separates into* labor, process, and product.

We can apply analytic distinctions between labor, process, and product to modern information retrieval activities and systems, with labor as human labor, process as a formalized process and often delegated to technology, and product as the outcome of labor and process (for instance, records or index descriptions).

We then introduced distinctions specific to mental labor. Mental labor could be semantic or syntactic, linking a widely held distinction between levels of analysis to mental labor. Semantic and syntactic mental labor could be regarded as separating from undifferentiated mental labor.

Human mental labor *separates into* semantic labor and syntactic labor.

A distinction then could be made between semantic and syntactic mental labor for both premodernity and modernity. Semantic mental labor corresponded to the work involved in conducting transformations on signs motivated by their meaning, and syntactic labor corresponded to transformations reduced to motivation from patterns. Semantic mental labor was irreducibly and directly human.

Semantic labor → human labor.

Expressing the relation between semantic labor and human labor as material implication allows for the possibility that human labor is more extensive than semantic labor, incorporating syntactic aspects.[2] In contrast to semantic labor, syntactic labor could be transferred to technology, and labor would become process.

Syntactic labor *can be transferred to* information technology (labor becomes process).

A technological process was *a priori* syntactic in character, operating on patterns and not directly on meaning. In premodern historical practice, syntactic labor was delegated to human clerical labor; under modern conditions, it is increasingly transferred to information technology as a machine process.

Transformations of Selection Labor

Selection power remained a relatively stable primitive or atomic fact as a quality of human consciousness; selection labor, although persistently present, modulated with historical transformations in information technology. In developing the proposition that selection power is produced by selection labor, we focused primarily on transformations of selection labor as a form of mental labor understood by the distinctions established within labor and mental labor.

When orality emerged into literacy, description labor in information systems included the cognitive labor of memory and recall as well as the bodily and communicative labor of public oral expression. Embodied in a socially designated individual, the Icelandic law-speaker exemplifies such an information system (Njal 1280/1960, 306–308). In oral speech, the process of production and the product (audible speech) disappeared with the process, leaving no trace outside the memory of the auditors; process

and product did not fully separate from labor. A distinction of syntactics from semantics could be recovered but is not strongly marked in oral societies.

For information systems under written literacy or premodernity, selection labor progressively separated into description and search labor, and search labor emerged as the work involved in traversing or searching the products of description labor.

Selection labor *separates into* description labor and search labor.

Analytically and substantively, selection labor represents the sum of description and search labor following separation. The historical separation of description and search labor from selection labor supports the proposition that selection labor can be distributed between description and searching. Separating products from labor introduces the possibility of distributing the products of description labor (catalog records, catalogs, and bibliographies); human labor congeals in those products. We can also delegate aspects of description processes to clerical human labor, with that labor assisted by technologies associated with writing as well as by special-purpose information machines that produce and sort written utterances.

Under modernity, the possibility and actuality of transferring syntactic labor to technology sharpens the distinction between semantic and syntactic labor for both description and search labor.

Description labor *separates into* description labor semantic and description labor syntactic.

Search labor *separates into* search labor semantic and search labor syntactic.

The semantic aspects of description and search labor, including the understanding of syntactic processes, remain irreducibly and directly human.

Description labor semantic → human labor.

Search labor semantic → human labor.

Previously conducted by direct human work, we can now transfer the syntactic aspects of description and search processes to information technologies, predominantly to the computer as a universal information machine programmed to function as an appropriate special-purpose information machine.

Description labor syntactic *can be transferred to* information technology (labor becomes process).

Search labor syntactic *can be transferred to* information technology (labor becomes process).

Thus, description and search labor are more fully revealed as both semantic and syntactic in character.

Summary

Our synthesis revealed a pattern of progressive development. Originally undifferentiated categories such as labor and selection labor separate into different aspects. The pattern revealed is essentially additive, congruent with a broader conception that human history involves the cumulative development of human capacities (Hobsbawm 1998, 41).

Figure 4.1 illustrates the central propositions underlying this discursive exposition as a sequence of summary statements [3]. The order of the main section in the discursive developments of the labor theoretic approach, both in the full expositions in the earlier chapters and recapitulated here, corresponds to the sequence of statements and can be mapped to it. The logical connective of material implication is used to relate concepts and activities distinguished, where it adds value to understandings obtainable from ordinary discourse. Material implication was classically the most difficult and productive of the logical connectives, and it emerges here as a highly significant crux, particularly illuminating the production of selection power by selection labor.

To aid intelligibility, we present these central propositions in a diagram restricted to the primitive terms, central proposition, and transformations of selection labor (see figure 4.2). The diagram also incorporates supporting propositions but does not show them explicitly. Selection power is given priority by its placement at the head of the diagram, where it rests on selection labor. The submerged position of selection labor is congruent with its congealing and disguise in its products and with its theoretical neglect. When exposed, selection labor significantly supports selection power. If the line indicating the distribution of selection labor (between description and searching) is understood as the beam of a balance or scales, selection labor must be considered as a hinged or potentially broken beam. For instance, if description labor is not helpful to the searcher,

Primitive terms

- Selection power
- Selection labor

Central proposition

- Selection power → Selection labor

Further propositions

- Labor includes mental labor

- Human mental labor *can be transferred to* Information technology
- Labor *separates into* Communal labor and Universal labor
- Labor *separates into* Labor, process, and product

- Human mental labor *separates into* Semantic labor and Syntactic labor
- Semantic labor → Human labor
- Syntactic labor *can be transferred to* Information Technology (labor becomes process)

Transformations of selection labor

- Selection labor *separates into* Description labor and Search labor

- Description labor *separates into* Description labor semantic and Description labor syntactic
- Description labor semantic → Human labor
- Description labor syntactic *can be transferred to* Information Technology (labor becomes process)

- Search labor *separates into* Search labor semantic and Search labor syntactic
- Search labor semantic → Human labor
- Search labor syntactic *can be transferred to* Information Technology (labor becomes process)

Figure 4.1
Summary of synthesis.

the quantity of description labor could increase without reducing search labor. The analogy of beam or scales remains valuable for giving iconic form to the idea that labor can be distributed between description and search.

In its current development, the argument is slightly static, and no dynamic compels the transformation of relations between categories from "can be" to "is."

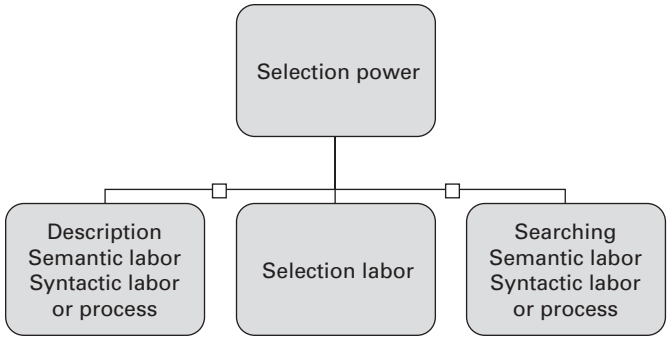


Figure 4.2
Representation of synthesis

Transformation into a Dynamic

A dynamism that compels the transfer of human syntactic labor to technology is found in the conditions of direct human labor, including its costs and elements of drudgery.[4] Discerning a widespread preference, particularly under capitalism, for reducing the costs of processes by abridging direct human labor is a specific and historically rooted position. It is distinguishable from Zipf's universalized principle of least effort or of resistance to labor (1936). The equivalence of automatic machine to slave labor (Wiener 1954, 162), a condition not often adopted voluntarily, should also be recalled. "Can be" relations between categories are transformed into "is."

Syntactic labor can be transferred to information technology (labor becomes process.)

is transformed into

Syntactic labor is transferred to information technology (labor becomes process).

The transfer would apply to both syntactic description and to syntactic search labor.

Description labor syntactic is transferred to information technology (labor becomes process).

Search labor syntactic is transferred to information technology labor (labor becomes process).

Thus, direct human labor primarily becomes semantic labor, and the transformation applies similarly to description and search labor— both aspects of selection labor.

The concepts of selection power and labor, and of semantic and syntactic labor and processes—all components of the model developed—have been abstracted out from historical and current information practice and from ordinary discourse, rather than imposed. In such a process of abstraction, there is an analogy with the construction of models of the computational process, not as ends themselves, but as aids to understanding. After grasping the fundamental matters, we can bring understanding, "which we could never obtain while immersed in inessential detail and distraction" (Minsky 1967, 2–3), back to the practical world. Automata theory develops models of the computational process partly as ends in themselves rather than as an endeavor that might inform practical understanding and has often become increasingly technical (Boolos and Jeffrey 1989). The challenge here is to show that the determining forces identified for information systems really do influence real-world systems.

Decision Practice

Although producers of major information systems would not formulate their policies and activities according to the concepts identified here, their activities are constrained and guided by the determining forces we have discussed. The effect of the determining forces can vary with the intentions and the producer's market.

A common and largely unvarying characteristic of the movements produced by motivating forces involves transferring direct human labor to technological processes. Under premodern conditions, the description processes (for instance, creation of indexes from records) were accomplished by human syntactic labor. Today those tasks are accomplished most frequently by technology, strongly implying a desire to avoid the costs of direct human labor. Description processes used previously to create certain description products in fixed media (for instance, different orders for indexes or records) are effectively moved from description to potential instantiation in searching, economizing on product creation and enabling a greater variety of orders.

In contrast to premodern practices, systems with syntactically based description processes—human labor transferred to technology—have proliferated, as exemplified by Internet search engines. Primarily syntactic and machine processes increasingly generate the descriptions used for searching the texts of works or citations between works on both Amazon.com (Amazon 2007) and Google.com (2007a, b, c). (Amazon's "search inside this book" facility offers a concordance for the particular text, and Google's descriptions or indexes provide weighted concordances; machine processes produce both systems syntactically). Therefore, human labor embodied in the written text of the documents described provides a resource for description. Description processes can aid specificity in retrieval, but the very number and diversity of results obtained may transfer selection labor to the searcher. The links or traces left by semantically guided explorations of resources can be exploited to determine the order of references in retrieval and could be regarded as products of a form of unpaid semantic description labor. The increasing dominance of the Google search engine, which might appear to contradict assertion of the proliferation of syntactically based systems, can be regarded as the diffusion of an effective syntactic system, with some monopolistic forces aiding its dominance. The market for information services has been regarded as acting as a regulative mechanism that eliminates or corrects errors; a better mechanism of correction remains undetermined (Swanson 1980, 128). Although syntactic processes for selecting and ordering documents may be open to unlimited variations derived from primitive computational operations, only some of these variations may be interesting or useful. One task for information retrieval research would involve attempting to understand the effectiveness of Google, including the effects of syntactic transformations on the semantic interpretation of words and phrases; subsequent chapters will address these issues.

Appropriately exploited for searching, the greatly increased scope of syntactically generated descriptions can enhance human capacities. For example, using Google's advanced book search function to find the exact phrase "power of knowledge objectified" yields a range of discussions of Marx's (1859/1973) conception of technology: "*organs of the human brain, created by the human hand*: the power of knowledge, objectified" (706) (Google 2007c). Such a search requires understanding patterns of

diffusion within the primary literature as well as the searching process. Regarding primary literature, Marx's phrase comes from the dominant English language translation of the *Grundrisse*, published in 1973 and not yet subject to processes of secondary diffusion that might have enabled discussion of Marx's theme of technology as a human construction without using the significant phrase. For searching, the understanding that phrase searching often exclusively yields tokens of the intended type could have been obtained by reiterated searches on different phrases.[5] Theoretically, we can explain the correspondence of tokens to type from the perspective of information theory regarding transition probabilities between letters and also between words, understood as cohesive groups of letters (Shannon 1951/1993). Chapter 7 discusses the understanding of the structure of written language obtainable from information theory. Semantic description processes bypassed in the discovery of documents still may be exploited to obtain copies of those documents from libraries or bookshops, using their own humanly and semantically constructed records. The enhancement of human capacities itself creates partly novel difficulties—selection from an enlarged body of potentially relevant and now more readily discoverable primary literature, and understanding how to exploit syntactically generated descriptions. Developments in description practices have increased and slightly transformed search labor.

Distribution of the products of semantic description labor (for instance, catalog records) emerged in premodernity and continues in modernity with elements of continuity and modulation. Thus, the costs of human semantic labor required to produce descriptions are effectively shared, although semantic labor itself is not directly divided. *WorldCat*, for instance, contains the products of semantic description labor as catalog records. Records are created by communal labor, guided by the universal labor embodied in codes, understandings, and precedents for document description. The products of communal labor (catalog records) are distributed to participating libraries, which use syntactic and primarily technology-based processes to incorporate them into their own institutional catalogs (WorldCat 2007). Amazon.com also uses products of human semantic description labor as records for books, although Amazon's standards are less full and exacting than those required by *WorldCat* (WorldCat 2007;

Amazon 2007). In contrast to the specificity obtainable from syntactically generated descriptions, deliberately human-assigned descriptions may offer generic power and control. For some systems (for instance, Internet search engines) semantic description labor used to assign metadata cannot necessarily be exploited directly and separately in searching. Distribution of description products reveals continuity with premodernity, and modulation emerges when machine processes supplant clerical labor to integrate products into intrainstitutional catalogs. We can characterize these practices as distribution of the products of human semantic description labor. The costs of that labor are thereby shared and we can understand the dominance of *WorldCat* in simple economic terms.

Inheritance of the absence of precise coding for oral speech and nonwritten graphic forms implies a limitation on the effective extension of established syntactic procedures to such forms. For searching and retrieval but not inspection of results, Google's image search relies primarily on descriptions generated from written verbal forms discovered in proximity to the images (Google 2007b). Other characteristics of computer-held images (for instance, type of file) amenable to automatic exploitation also can be searched. Retrieved results are presented iconically; the reduced size of the display images enables further scanning. The contrast between the treatment of symbolic and iconic graphic signs implies a connection between labor and epistemology, where written language embodies accumulated or universal human encoding labor (for instance, standardized orthography and clearly marked boundaries between words).

Box 4.1
Difficulty of Analyzing or Segmenting Oral Speech

> Here, written communications present one kind of problem and voice communications another—generally considered more difficult. To sort out the second type requires a "word-spotting" capacity—essentially a computer program that can distinguish between spoken words in a multitude of languages . . . Admiral Bobby Ray Inman, a longtime director of the American codebreaking and Comint efforts as chief of the National Security Agency, once admitted in public that word spotting for voice systems remained a dream. "I have wasted more US taxpayers' dollars trying to do that," he said, "than [on] anything else in my intelligence career." (Powers 2005, 22)

The determining forces isolated by theoretical discussion have emerged in major information systems as highly significant influences on real-world developments. Technological processes are *a priori* syntactic in character. The most compelling aspect of the dynamic observed, determined by relative costs, has been the transfer of syntactic processes from direct human labor to humanly constructed and adopted technology. Although not necessarily applied widely, human semantic description labor persists both as an inheritance and for its current value. Where semantic description labor had value in enhancing selection power and reducing search labor, its absence tends to transfer selection labor to the searcher. One writer has suggested that a theory of bibliographic searching may replace a theory of bibliographic description (Wilson 2001). A task for information retrieval research might involve expertise and strategies that can assist exploitation of syntactically generated descriptions. Once understood, strategies may not be complex: consider, for instance, the contrasting results characteristically obtained from word and phrase searching.

The forces identified in the model contain no intrinsic dynamism that compels transferring human semantic selection labor from description to searching, but they may have that effect in many contexts. In a relatively closed and controlled system, the fact of consumption of material may be assured and patterns of consumption—one epistemological basis for the choice of mode of description—can be anticipated (Shera 1952/1965; 1961). Even then, assigning objects to categories can be problematic and highly inconsistent. More open systems cannot assure the fact of consumption, and patterns of consumption are difficult to anticipate: "if you watch it [display of Google queries] long enough, the different queries show how diverse the world is" (Weisman 2002). Thus, practices that reduce human labor in description and transfer work to searching may proliferate.

Decision Variation

Decision practices reveal evident commonalities, variation, and an underlying value. Evident commonality lies in transferring established description processes to technology. Variation occurs primarily in two loci: the extent of syntactic description processes applied—what proportion of the original document to capture and represent—and the presence or absence

of the products of semantic description labor and the type of product. Some level of human semantic search labor is inescapable. An underlying value, embodied in real-world systems and influencing practice, is selection power.

Different aspects of variation have different costs and effects. Although not without cost, expanding the scope of syntactic description processes has limited costs and may enhance human capacities. In contrast, semantic description labor is costly. The presence of the products of semantic description labor may enhance selection power and reduce search labor and their absence can limit selection power and incur high levels of search labor.

Decision Considerations

We can construct decision considerations from the contrasting costs and effects of the two aspects of variation. The low costs of syntactic description processes and the possible enhancement of human capacities favor their instantiation. Restrictions on the scope of machine description processes may be only a transitional stage, with developments inhibited by inherited beliefs and attitudes, particularly in library rather than broader information contexts. The primary consideration in making decisions concerns semantic description labor and the associated distribution of human semantic labor between description and searching. Because description labor is more amenable to deliberate control than search labor, the specific consideration would cascade from deciding whether to apply semantic description labor at all, to determining which type (including exhaustivity of description), to resolving whether to draw upon any existing products of semantic description labor. Therefore, decision considerations relate primarily and directly to human labor, secondarily to machine processes, and only slightly to the materials consumed in the products of labor and processes.

Value of a Labor Theoretic Approach

We identified a historical dynamism for information practice that distinguishes modes of information: primarily oral, written literate (premodern), and computer technologies (modern). Excepting the recently

dominant but currently eroding information retrieval research tradition, the value of selection power was relatively consistent. In contrast, selection labor emerged experientially and as an empirically observable practice (written literacy), then divided further into description and search labor, before reconverging as a single substantive category. Historical and empirical considerations reinforced the validity of the proposition that selection power is produced by selection labor, most clearly when information has exosomatic and material aspects. Further consideration of the value of the labor theoretic approach should include:

- its formal qualities as a theory,
- its relation to the objects described,
- its connection to ordinary discourse and common experience, and
- its absorption of preexisting theories.

Formally, the theory has simplicity and economy and can be reduced to a brief series of statements—qualities usually valued strongly in a theory. In relation to objects described, the theory adopted an expansive understanding of information retrieval systems, accompanied by ostensive exemplification rather than restrictive definition. The theory was able to comprehend systems in oral, written (premodern), and computational (modern) modes. We identified a dynamic of change and possibilities for deliberate and informed intervention in modern systems. Therefore, we can consider the theory as comprehensive and powerful in relation to the activities it takes as its object. The combination of simplicity and economy with comprehensiveness and power makes the theory parsimonious, possibly representing a final reduction to essential and inescapable elements.

The comprehension of real-world practice is matched by closeness to ordinary discourse and common experience. The concepts of selection power and selection labor and the distinction of semantic from syntactic labor have analogues in ordinary discourse conceptions and everyday information practices. For instance, the analytical distinctions introduced and employed between labor, process, and product all emerge from real historical transformations of lived categories. Rooting in ordinary experiences should give the theory further qualities of robustness and wide applicability.

We have selectively absorbed preexisting theories from librarianship, indexing, and classic information-retrieval research and also absorbed a

concept from information society discussions. Selection power was taken from librarianship and indexing, but the preference of those activities and disciplines for human description was transformed into a fuller understanding of the distribution of human semantic labor between description and searching, and recognition of the possibility of transferring forms of labor to technology. Query transformation—implicitly valued in classical information-retrieval research—was incorporated into selection power both practically and theoretically as a special case, similarly produced by selection labor. Informational labor, recognized in information society discussions, was differentiated into semantic and syntactic mental labor, and human labor was distinguished from machine processes. Technology has not been repressed but rather has been incorporated explicitly into the theory. Therefore, we have achieved a synthesis of preexisting theories that adopts their elements of value and discards aspects that obstruct understanding.

The relation between the labor theoretic approach and preexisting theories has a partial and revealing analogy to revolutions in mathematics. Classically, revolutions in mathematics have changed interpretation (verbal metalanguage) while retaining form (symbolic object-language), thus economizing on the accumulated intellectual labor (universal labor) embodied in known transformations in object-language:

> And in thus preserving the form while modifying the interpretation, I am following the great school of mathematical logicians who, in virtue of a series of startling definition, have saved mathematics from the sceptics, and provided a rigid demonstration of its propositions. (Ramsey 1925/1990, 219)

In this instance, we have brought the interpretation into rigorous accord with essential or inescapable components of form or practice, particularly through the idea of selection. Selection inherent in the process of retrieval is matched by the value placed upon selection, and the addition of power as the fundamental value for information retrieval. Therefore, the value of selection power is congruent with the process of selection.

The theory then meets a rigorous test for knowledge, as "an ideal reproduction of the external world serviceable for cooperative action thereon" (Childe 1956, 54). The theory has elements of idealization in its abstraction out of common activities from empirical reality, but the abstraction gives enhanced understanding of those activities (for instance, the distinction between semantic and syntactic mental labor and their relation to

technology). Distinctions established have been returned to the external world, yielding an enhanced understanding of major information systems and identifying a dynamic for change. The theory also serves cooperative action, enabling guided and deliberate intervention in the dynamic (for instance, the distribution of semantic labor between description and searching) rather than simply allowing reproduction of the dynamic that is less than fully conscious.

Just as the owl of Minerva takes flight at evening, understanding grows toward the end of a process. In this instance, human description labor has been fully identified as a category complemented and challenged by syntactic machine description processes. For future development, particular aspects of the labor theoretic approach may obtain special significance. For instance, the enhancement of human capacities enabled by the increased scope of description processes also raises the issue of how best to determine the most effective ways of exploiting the descriptions enabled. Existing research themes could be transformed and carried forward. For instance, recognition of the multifaceted nature of relevance, previously understood to require different methods of human description (Wilson 1973), now could be conceived as how to exploit syntactically generated descriptions along the dimensions of relevance identified. Crucially, the labor theoretic approach can absorb possible future developments, and fuller empirical richness is possible within its established conceptual framework.

As computational technologies become increasingly diffused there is a dissolution of the gestalt of the computer (Rosenberg 1974), in the particular sense of extravagant expectations of the computational process and of the transfer of human intelligence to technology. The implicit transfer of human judgment to technological processes in query transformation may have partly motivated the growth and diffusion of the classical information retrieval tradition. The experience of the practice of computation is reducing the contrast between mystical practice—in the sense of proposals for processes—and restrained theory; the theory of computation itself incorporates ordinary discourse or natural conceptions of the computable. The revolution in the mechanization of mental labor (Minsky 1967, 2) can now be seen as a late-twentieth-century revolution that has stabilized at the syntactic level of computational transformations and the interface to systems.

Conclusion

At the semantic level, the process of information retrieval may remain far more enduringly intractable and, despite technological and system developments, not amenable to teleological transformation. One mid-eighteenth-century comment, by a lexicographer who relied partly on his own accumulated knowledge for making a monolingual dictionary (Boswell 1791/1980, 131–132),[6] still gives the most convincing account of the process of information retrieval:

> When first I engaged in this work, I resolved to leave neither words nor things unexamined, and pleased myself with a prospect of the hours which I should revel away in feasts of literature, the obscure recesses of northern learning, which I should enter and ransack, the treasures with which I expected every search into those neglected mines to reward my labour, and the triumph with which I should display my acquisitions to mankind. . . . But these were the dreams of a poet doomed at last to wake a lexicographer. I soon found that it is too late to look for instruments, when the work calls for execution, and that whatever abilities I had brought to my task, with those I must finally perform it. To deliberate whenever I doubted, to enquire whenever I was ignorant, would have protracted the undertaking without end, and, perhaps, without much improvement; for I did not find by my first experiments, that what I had not of my own was easily to be obtained: I saw that one enquiry only gave occasion to another, that book referred to book, that to search was not always to find, and to find was not always to be informed; and that thus to persue perfection, was, like the first inhabitants of Arcadia, to chace the sun, which, when they had reached the hill where he seemed to rest, was still beheld at the same distance from them.
>
> I then contracted my design, determining to confide in myself, and no longer to solicit auxiliaries, which produced more incumbrance than assistance: by this I obtained at least one advantage, that I set limits to my work, which would in time be finished, though not completed. (Johnson 1755/1982, 21–22)

Mental labor is both made explicit and alluded to in repeated analogies between literature searching and physical mining. The Aristotelian notion of deliberation—"to deliberate whenever I doubted"—is also critiqued implicitly. The contrast between "finished" and "completed," a subtle distinction between the meanings of words made in a lexicographic context, has relevance here: this particular sequence of chapters ends, but the research agenda is not complete.

5

Retrieval from Full Text

Introduction

The remaining chapters focus on retrieval from full text, which is understood as a further transformation of selection labor enabled by modern information technologies; particular aspects of selection labor identified by the labor theoretic approach rise to prominence. Specifically, the quantitative expansion in the scope of description process has transformed the possibilities for search labor. Description processes remain syntactic and machine-based, while search labor includes directly semantic elements (for instance, the choice of words for search) and semantic understanding of syntactic components (for instance, the logical combination of search terms and understanding of the effects of description processes on the meanings attached to words). This chapter proceeds from the concrete to the abstract, from examples of retrieval from full text to discussions informed by scholarly sources accepted as relevant to understanding the semantics of written language, particularly regarding the idea of language as a nomenclature and the insufficiency of this conception of language. The relation of the labor theoretic approach to full text retrieval is precisely specified. Subsequent chapters develop a fuller and more positive understanding of the semantics of written language (understood as the transformations of meaning in the processes of description and searching) and of syntax (understood as patterns of difference and replication in words and sequences of words).

Examples

Unquiet cooccurrences

Quotidian metamorphise

Happiest legumens

The fragment shown above represents a novel literary form in which each line combines a unique co-occurrence on Google (2004).[1] The primary explicit focus of composition centers on the combinatorial possibilities of the signifier or expression, although the signified or meaning reemerges forcefully in thematic resonances containing unique combination and transformation of terms. Composition in this instance also involved a more deliberate search for a pun that would allude to *hapax legomenon*, the scholarly term for a word that occurs only once in a classical or biblical corpus. Because the implied Boolean AND characteristically retrieves a number of documents from a large corpus, discovering unique co-occurrences—even those devoid of semantic interest—is difficult.

The contrast between the retrieval results expected from combining individual search terms with AND and OR would be relatively well-known and understood, both from the historical experience from the 1970s of searching systems based primarily but not exclusively on the representation of single terms taken from source documents, and also from the theory of computation. Resemblance in the initial ordering of references would partly disguise the empirical differences in the retrieved sets obtainable from search engines using a phrase rather than an all the word search, but the differences would be discernible by the number of documents retrieved.

Experience in searching newspaper databases—regarded as more ordinary than scientific or technical discourse—and as anticipating the issues encountered when other forms of ordinary written discourse would become subject to computer-based retrieval also revealed the multivalency—multiple values or meanings, that single words acquire when detached from their syntagmas. Retrieval recovers the syntagma and reveals the variety of syntagmatic occurrences. For instance, searching for the topic *financing of university libraries* in 1993 with a query equivalent to *university AND library AND finance* retrieved unintended documents:

THE GUARDIAN Copyright (C) Guardian Newspapers Ltd,
1984,1985,1986,1987,1988,1989,1990,1991,1992,1993

>get university
GET UNIVERSITY
3678 ITEMS RETRIEVED SINCE 1FEB92

>pick library

PICK LIBRARY
143 ITEMS RETRIEVED SINCE 1FEB92

>pick finance

PICK FINANCE
8 ITEMS RETRIEVED SINCE 1FEB92

>context 1

CONTEXT 1

1....
GDN 23 Jan 93 Latest Positions: Sex by Vatsyayana (1770)
....
Kama is love, Sutra is art. The infamous treatise on the art of love was part of a trio of works; the others were the Artha Shastra, one of the world's earliest manuals of statecraft, written by Kautilya, chief minister of Chandragupta Maurya, and the Dharma Shastra. . . . Then as now, there was no romance without FINANCE.
....
Later, he became a professor at the Hindu UNIVERSITY of Benares and a director of the college of music. In 1954, he moved to Madras to become the director of the Adyar LIBRARY.

None of the query terms received in the retrieved passage are used in senses far removed from their ordinary discourse denotations and connotations. Other retrieved results suggested a university as a site for transition from adolescence to adulthood as well as a place of learning—a more formal sense consistent with discussions of aboutness in information retrieval research. Similarly but not precisely parallel, a search for public discussion of the Research Assessment Exercise (RAE)—the evaluation of publicly funded research in the United Kingdom—in 1997 supplemented the phrase *research assessment exercise* with the acronym *RAE* combined with *research* to retrieve documents that did not mention the full phrase (Warner 1997, 264–265).

(Research Assessment Exercise) OR (RAE AND research)
Did Englishmen eat each other? In the 150 anniversary year of Sir John Franklin's death, Tom Pow considers why his failed endeavour overshadowed the greater achievement of the Orcadian explorer John RAE . . . . Anyone who has visited the Scott Polar RESEARCH Institute and seen the final copperplate letters, written with all hope gone, with breath averted so paper would not freeze. (Pow 1997)

The logical operators used in searching the newspaper databases continue to Internet search engines, although expressed in different ways. Some learning from experience is revealed in the use of phrase searching.

In contrast to combining search terms by AND, phrase searching tends to recall a limited number of documents that contain thematic or semantic similarities to the intended subject or signified in the query. For instance, searching for the phrase *direct semantic ratification*[2] recalls a restricted number of documents on the topic of direct semantic ratification, often concerned with contrasts between orality and literacy and beginning to address issues concerned with electronic communication (table 5.1). Similarly, searching for the phrase *Jack Kennedy was a friend of mine* retrieves a limited set of documents concerned with or deliberately alluding to the vice-presidential debate between Lloyd Bentsen and Dan Quayle. The discussions recalled seldom noticed the rhetorical effectiveness of the contrast between the deliberate informality (*Jack Kennedy*, implying the personal friendship claimed) and the formality of *Senator* (alluding to Quayle's official role) in Bentsen's continuation, *Senator, you're no Jack Kennedy* (Google 2004). The variable verbal forms given for other parts of Bentsen's remarks could be ascribed to the inexactness of unchecked verbal memory.[3] Thus, phrase searching is used as a compensatory mechanism for the mutivalency of separate words. A recent innovation by Amazon.com, which displays statistically improbable phrases from the text of documents and uses them to make interdocument connections—reveals similar practical rather than necessarily theoretical understanding (Amazon.com 2005).

Table 5.1
Effects of different search strings

Search statement	Retrieved items
"direct semantic rarification"	c.83
direct AND semantic AND ratification	c.8700
direct OR semantic OR ratification	c.28000K

Source: Google 2004.

Practical Understanding and Theoretical Knowledge

Understanding theoretically—rather than just experiencing practically or exploiting—the differences between term and phrase searching requires movement beyond the theory of computation to linguistics, specifically to Saussure for an account of signification in written language (Saussure 1916/1983), and to information theory for the possibilities and constraints on the combination of words into phrases (Shannon 1948/1993). Comprehending patterns revealed by phrase searching might illuminate why search engines have been perceived as effective. For instance, retrieving documents exclusively on a topic corresponds to a principle intended but not always achieved purpose of humanly assigned index or metalanguages and to an aim of classical information retrieval, particularly when oriented toward precision. With phrase searching, some expertise and mental labor has been transferred from description to searching. Synthesizing aspects of the seldom-connected fields of Saussurean linguistics and information theory could be facilitated by establishing analogies between their principal concepts: between the syntagma (the linear sequence of utterance) from linguistics and the message from information theory and the paradigm (the network of associations) and the messages for selection.

This reverses a commonly assumed relation between theoretical knowledge, practical experience, and understanding and gives practical understanding priority over theory. Philosophical antecedents to this reversal are found in Marx's insistence on the value of rising from the abstract to the concrete (1858/1973, 101) and in Vico's preference for practical understanding over philosophical universals (1710/1988, 62). Giving priority to practical understanding may be particularly appropriate for human communication, since the ability to communicate characteristically precedes analysis of communication. Vico's comments on his own account of the transition from primary orality to literacy, regarded as the coevolution of spoken and written language, are particularly apposite:

> All this seems more reasonable than what Julius Caesar Scaliger and Francisco Sánchez have said with regard to the Latin language, reasoning from the principles of Aristotle, as if the peoples that invented the languages must first have gone to school with him! (1744/1976, 153 § 455)

Similarly, practical understanding may have preceded theoretical analysis regarding the transition from paper-written to electronic literacy.

The additional value of theory lies in the enhanced understanding of practice and in its comprehension and generalization of experientially and empirically obtained understandings. The choice of particular theories (for instance, Saussurean over Chomskyan linguistics) can be supported by experience, with a dialectic relation between practice and choice of theory. In this instance, some particularly strong theories are employed. Mediated by semiotics, Saussurean linguistics provides the most sophisticated and powerful method known for analyzing signification (Warner 1994, 9–15). Information theory (Shannon 1948/1993) sets limits for the transmission of messages of comparable standing and significance to the limits established for computation by automata theory (Minsky 1967).

The strength of theories is such that they may assist in establishing that the emerging elements of stability in practice correspond to a teleological statue or plateau that only could be transcended by disproving the theories. Generally recognized as particularly strong, automata theory established fundamental possibilities and limits for computation (Cockshott and Michaelson 2007). Information theory gives a model of communication that identifies the fundamental entities required for signal transmission. Information theory has not been disproven, although understanding of its legitimate application to message and signal transmission has been obscured by its analogical interpretation to include semantic understanding (Tidline 2004). Saussurean linguistics can yield a very powerful, although not previously developed, account of transformation for meaning for full-text description and searching. From the perspective of automata theory, failure to progress in practical applications despite repeated efforts may strongly imply that processes are computationally intractable. The theories here may reveal the sources of the continuing intractability of information retrieval, in both its ordinary discourse and its technical sense, and even identify the sources of possible noncomputability that may underlie its intractability. Contemplating moving beyond the current, relatively stable elements of practice might then require a refutation of firmly embedded and not-yet-disproved theories. In particular, human semantic labor, possibly invested in search rather than description, may remain inescapable to counter the intractability of the process of retrieval.

This chapter adopts a rigorous understanding of knowledge and theory as knowledge. One powerful understanding of knowledge is that it must be communicable and useful, amenable to translation into successful

action. Knowledge continues to be understood as "an ideal reproduction of the external world serviceable for cooperative action thereon" (Childe 1956, 54). A distinction of knowledge from information, independently developed in subsequent literature (Blair 2002, 1020–1021), is implied. Childe's "ideal reproduction" is understood to include a degree of idealization and also the active participation of receivers in reproducing knowledge. In this context, idealization is realized in the abstraction and simplification of elements of language and message transmission; active participation and the useful nature of knowledge emerge in the utility of the abstractions chosen for informing subsequent use and design of information retrieval systems. The knowledge embodied in information technology is acknowledged but treated as a given and as a basis for the development of further knowledge. The embodied knowledge is received in objectified form, as material technology (Warner 2004, 5–35).

Conceptions of Language

The conception of language implicit in the searches recounted above was concisely articulated by Saussure in his *Course in General Linguistics*, first published in 1916—far in advance of the possibilities of discovering the use of words offered by modern technologies for information retrieval (1916/1983, 65).

For some people, a language reduced to its essentials, is a nomenclature: a list of terms corresponding to list of things. For example, Latin would be represented as:

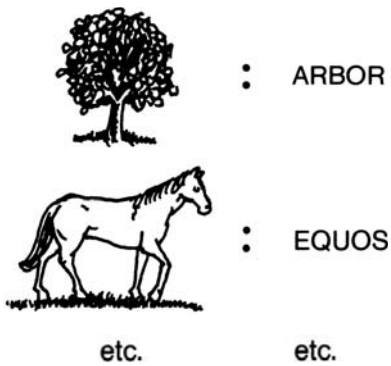


Figure 5.1
Source: Saussure 1916/1983, 65.

In one instance, *library* was conceived primarily as referring to a physical place for information-related expenditures. Similarly, *university* was understood spatially, and *finance* represented a definable, related activity. Saussure critiqued the view of language as a nomenclature, arguing that a word obtains its value from a "network of forever negative differences" (Culler 1988, 52). Experience with information retrieval systems has tended to confirm Saussure's proleptic insight, particularly regarding his view that we cannot understand language fully as a nomenclature.

We can trace the persistence of the ordinary discourse notion of language as a nomenclature, possibly experientially eroded now by modern information retrieval, to inculcation into written literacy. In Saussure's exemplification of this conception of language, equivalence is indicated between words and pictures of objects, corresponding to a child's alphabetic primer. The relation between word and image is known to break down when tested by more abstract entities— how could the Research Assessment Exercise be drawn?—but we may still assume a similar relation of equivalence between word and concept. In a similar way to Saussure's exemplification of language as a nomenclature, Augustine's description of the acquisition of language implies a relation of equivalence between spoken sounds and visually perceived objects:

> I can remember that time, and later on I realized that I had learnt to speak. It was not my elders who showed the words by some set system of instruction, in the way that they taught me to read not long afterwards; but, instead, I taught myself by using the intelligence which you, my God, gave to me. For when I tried to express my meaning by crying out and making various sounds and movements, so that my wishes should be obeyed, I found that I could not convey all that I meant or make myself understood by everyone whom I wished to understand me. So my memory prompted me. I noticed that people would name some object and then turn towards whatever it was that they had named. I watched them and understood that the sound they made when they wanted to indicate that particular thing was the name which they gave to it, and their actions clearly showed what they meant, for there is a kind of universal language, consisting of the expressions of the face and eyes, gestures and tones of voice, which can show whether a person means to ask for something and get it, or refuse it and have nothing to do with it. So, by hearing words arranged in various phrases and constantly repeated, I gradually pieced together what they stood for, and when my tongue had mastered the pronunciation, I began to express my wishes by means of them. In this way I made my wants known to my family and they made theirs known to me, and I took a further step into the stormy life of human society, although I was still subject to the authority of my parents and the will of my elders. (498/1961, 29 § I.8)

Box 5.1
Multivalency of Words

"When *I* use a word," Humpty Dumpty said, in a rather scornful tone, "it means just what I choose it to mean—neither more nor less."

"The question is," said Alice, "whether you *can* make words mean so many different things."

"The question is," said Humpty Dumpty, "which is to be master—that's all."

—Lewis Carroll. *Through the Looking Glass, and What Alice Found There.* (Carroll 1871/1998, 183–186)

Augustine focused explicitly on the acquisition of oral language, but we can discern some influence from written literacy, particularly in the word to object relation. Words are not necessarily distinguishable units of oral utterance, in their aspect as signals, even when those utterances are influenced by written language. The culturally significant artifact of the monolingual dictionary both reinforces and undercuts the concept of language as nomenclature. Rather, it reinforces the concept by representing relations of equivalence between defined words and their definitions, although characteristically between verbal forms rather than word and image; it also undercuts the concept, particularly in more extensive dictionaries, by revealing the variety of signification that can attach to defined terms. Even extensive and historical dictionaries might still imply stable synchronic states of a language, showing clear transitions between historical uses of words and sharply differentiated, definable senses while assigning uses to specific single categories.

Construction of an equivalent to a nomenclature—a deliberately made language—has been a recurrent response to the perceived inconsistency of the relation between word and object or meaning in ordinary as well as scholarly and scientific discourse. In philosophy, Wilkins proposed a real character—intended to "signifie things, and not words"—that aimed to achieve clarity in argumentation (1668/1968, 21). In indexing, controlled vocabularies are intended as terms with stable meanings that label concepts more consistently than the language of discourse. Less commonly, a search for a nomenclature has become embedded in the language of discourse. A common but not universal assumption in the construction of a nomenclature argues that concepts and objects can be known independently of language and that agreement exists regarding fundamental notions. The excluded multivalency and indeterminacy of ordinary discourse tends to reemerge in the complexity of deliberately constructed language and also in the difficulty of assigning objects to categories. The theoretical sophistication of discussions about deliberately constructed languages has varied (Gardin 1973, 142), particularly regarding the granularity of isolation of concepts and the relation of concepts isolated to the language of discourse.

Retrospect and Prospect

We can distinguish a retrospect and a prospect regarding full-text retrieval. The retrospect comprises the theoretical development to this point—the labor theoretic approach and the further sources adduced in this chapter. In contrast, the prospect involves developing a fuller understanding of the emerging qualitative changes isolated by the retrospect, not just accounting for their existence. It will be constituted by the subsequent chapters.

Retrospect

The labor theoretic approach can explain the existence of full-text retrieval as a real-world practice. Its emergence can be understood as a product of the desire to exploit the enhanced selection power made possible by the quantitatively increased storage and processing capacities of modern, humanly constructed information technologies. Selection power continues to be implicitly and strongly valued, and it remains produced by selection labor, with aspects of that labor (particularly description labor) transferred to machine processes. Therefore, the value implicitly attached to selection power and its production by selection labor reveals consistency with the labor theoretic approach. The significance of selection power (and the value attached to it) and the inescapability of selection labor are further confirmed by systems that offer enhanced selection power but require forms of selection labor—systems currently entering widespread existence and use.

The underlying dynamism for change can be identified precisely. Change has emerged principally from the reduced costs of syntactic machine processes and the increased amount of resources that can be incorporated economically into those processes. The labor theoretic approach and the theories in which it is embedded can explain how substitution of machine processes for direct human labor reduces the costs of processes. The enhancement of human capacities that emerges from repeated substitutions is also explicable, if less frequently recognized, particularly by the labor theory of value. Therefore, the immediate origins for change are consistent with a broader emphasis on the significance of changes in the material basis of being or existence to practice, and with a more specific view of recent developments in information technologies as a revolution in the mechanization of mental labor.

We can locate the dynamism for change in relation to the categories offered by the labor theoretic approach. The primary location for immediate change lies in the intersection between syntactic machine processes and the description of objects (lower left section, table 5.2). The enhanced scope for machine description of objects obtained has had secondary effects on other aspects of information retrieval. The common source for change has been mediated by contrasting and sometimes antagonistic cultures of library and information science and Internet search engines, resulting in different secondary effects and practices.

For library and information cultures, the dynamic has flowed predominantly in two directions: toward the intersection between human semantic labor and the making and distribution of description products, and also toward invoking machine syntactic processes in searching (represented by the upper left quarter and the transition between the upper and lower right quarters of table 5.2). Producers of library and information resources have been reluctant to exploit the potentialities of describing the full-text of documents, although this was identified earlier as a possibly transitional stage that was no longer compelled by the storage or processing constraints of technologies. Therefore, only limited excursion has occurred into human semantic search labor, both in practice and in

Table 5.2 Distribution of human labor and machine processes between description and searching under modernity

	Description	**Searching**
Human semantic labor	Continuity and modulation (distribution of products of semantic description labor). Generic capacity.	Consequential (from increased scope of machine syntactic description practices) increase in possibilities for exploitation. Need to understand production of meaning and patterns of occurrence, in order humanly to exploit.
Machine syntactic process	Increase in scope of resources covered. Strong continuities in description techniques.	Invoking of machine processes (copying and sorting) rather than direct human syntactic labor/without syntactic labor conducted.

theory. Experimental information retrieval, which developed partly from library and information cultures, has had a marginally direct effect on practice and correspondingly limited secondary real-world effects flowing from the primary dynamism. The changes resulting from the mediation of the dynamics for change within the library and information cultures—the distribution of products of description labor and the invoking of machine processes for ordering results—are historically embedded and can be regarded as a quantitative expansion of existing practices.

For Internet search engines, change has occurred principally but not exclusively within the intersection between machine syntactic processes and description. Quantitative expansion in the potential scope of syntactic description progress has created some qualitatively significant but poorly understood changes in human semantic search labor. A consequential and highly significant dynamism flows toward the intersection of human semantic labor and searching, increasing the possibilities for exploiting resources (the change occurs in the lower left quarter of table 5.2, and consequential effects appear in the upper right quarter). From the systemic perspective developed in studies of the role of information services, changes in one aspect of a system that have unintended consequences in another aspect would be understood as a familiar phenomenon (Roberts 1997). Semantic description labor and the more formalized invoking of machine processes for search ordering and combination developed in and associated with the cultural milieu of library and information science are usually bypassed, but not always. The potential for exploitation of full-text description is much less embedded historically and less fully understood than developments in library and information science. It represents a qualitative change, not just quantitative expansion.

The persistence of underlying cultures and their influence on practice illustrates that practice and particularly theoretical consciousness can be underdetermined by the material basis of being: "[t]he tradition of the dead generations weighs like a nightmare on the minds of the living" (Marx 1852/1973, 146). The material basis of being, mediated by technology, directly affects practice in the sense of methods for constructing working systems and practical understanding of how to exploit those systems. In contrast, theory can be slow to change and dominated by its inheritance from previous modes of thought. Thus, practice can precede theory.

Studying practice can yield a highly specific focus for theoretical understanding. The elements of practice can be arranged along a continuum that ranges from contingent to inescapable. Contingent elements are understood as those that vary across contrasting cultures and practices. Conversely, inescapable elements strongly influence both library and information and Internet cultures and practices. Isolating the inescapable—and attempting to understand it—may yield the basis for a synthesis of the contrasting cultures. Any such synthesis likely will require a painful effort of deliberate articulation; even then, synthesis will remain incomplete at this stage.

Contingent Elements

It is possible to group significant contingent or variable elements into meaningful categories that reveal underlying analogies between contrasting cultures. Human description labor and its strongly semantic elements appear in the library and information sphere. The products of that labor—bibliographic records and the distinctions embodied in different fields of those records—can be exploited directly in searching. In addition, human description labor is highly and unquestioningly valued in theoretical discussions. Documents discovered by Internet search engines may contain humanly assigned metadata, but it is not necessarily available for direct exploitation. A significant development, both analogous and contrasting to human description labor, involves exploitation of the traces left by semantically motivated human searching, which produces effects analogous to those obtained from description labor (for instance, grouping of material).

Generic capacity—grouping material—occurs in both cultures but it is produced by contrasting methods. Library and information practices characteristically offer generic capacity by incorporating the products of human description labor (for instance, subject descriptors or canonical forms for authors' names). Evidenced by Google Scholar (2008), a contrasting method involves a computational collation of an article's variants through similarities in pattern. The sense of generic capacity common to both cultures is the minimal sense of gathering together; the more elaborate, humanly assigned structures are developed more frequently by library and information practices. The ordering of results represents one form of gathering, and practices in both cultures offer intelligible—or at

least explicable—orderings and discernible analogies. For instance, ranking results by the number of links to Internet pages may be analogous to ordering records of books by the number of libraries that hold them, or ordering journal articles by the number of citations received: the various orderings all function as surrogate indicators of diffusion or adoption.

Variable practices are intended to confer value, and the persistence of some variable elements in practice across contrasting cultures implies the possibility of value or significance transcending their particular method of instantiation, whether by human semantic description, search labor, or computational transformations. The market effect continues to function as a selection mechanism, progressively differentiating synchronic value from inherited practice. The particular forms of value conferred might be better understood by contrasting them with inescapable elements, laying bare the real reasons for their continued existence.

Inescapable Elements

Selection operations are inescapable and intrinsic to the understanding of information retrieval we have developed here; the particular selection operations may vary, although more in theory than in real-world practice. Experimental information retrieval began by critiquing the effectiveness of Boolean searching, and developed forms of retrieval it characterized as non-Boolean. However, directly Boolean operations are very common in real-world practice for the selection but not necessarily the ordering of retrieved documents and records. They also correspond to fundamental computational operations. Contrary to criticisms made in experimental retrieval, and possibly attracted by the transparency and intelligibility offered by Boolean operations, the dominant market effect is toward selection of Boolean systems. The crucial dimension for further analysis will involve abstracting the word or other segments of utterance from writing, which is realized in Boolean searching but may not be confined to it. Therefore, we can assume Boolean operations for our subsequent analysis, ensuring its relevance to real-world practice without significantly restricting its theoretical scope.

Inescapable elements emerge as the residue left behind following the separation of variable elements. Under modernity, selection operations are intrinsic to information retrieval, and all these operations can be generated from the fundamental computational operations necessary for

writing, erasing, and substituting symbols. Selection operates on data and inevitably encounters semantics and syntax—meanings and patterns—of that data or medium. Thus, both semantics and syntax are also inescapable. When the selection medium is written language, semantics undergoes a further translation into a more specific concern regarding production of meaning from written language and syntax becomes issues surrounding the replication and difference of words and sequences of words. These inescapable elements do not simply coexist but engage dialectically: computational procedures for selection that correspond to modernity interact with the production of meaning and patterns of replication and difference that embody strong inheritances from orality and from written literacy, or premodernity.

The dialectical interaction of computational procedures for selection with the semantics and syntax of written language—understood as patterns—is most fully and directly realized in Internet search engines, although more in practice than in consciousness or theory. The most distinctive, novel, and strongly emerging aspects of Internet search engines is not full-text retrieval itself—consider Biblical concordances—or the underlying processes that support selection, but rather the scope and extent of data covered, with variety merging into global heterogeneity. Enabled by the transfer of description process from direct human labor to machine and also by the reduced constraints of machine storage limitations, a quantitative change yields qualitative effects. In addition, search engines may be approaching a teleological state or plateau and may embody the possibilities approached by other systems. Systems that employ semantic description labor directly, and humanly, encounter patterns of production of meaning from written language, along with a subsequent dialectic between the descriptions and computational procedures for selection. Therefore, inescapable elements transcend the divide between the library and information and Internet cultures, although they occur differentially in the more novel, interesting, and possibly historically inevitable occurrences embodied in Internet search engines.

Contingent and Inescapable Elements

By separating the inescapable from the variable or contingent, we can now progressively isolate the set of activities that should form the focus for our prospective attention. For analytical purposes, we begin by

excluding highly variable elements: human description labor and practices that depend on the presence of the products of human description labor (including exerting generic capacity and certain forms of ordering) from the library and information sphere, and semantic search labor adapted to description, syntactically produced generic capacity, and the orderings of records from the Internet sphere. We can still include one slightly variable element, Boolean logic for selection, while recognizing the possibility of other computational selection procedures and also recognizing that the critical factor is abstraction of the segment of utterance from the line of writing, not Boolean logic itself. We can then focus our attention on the interaction of the inescapable elements; Internet search engines currently offer the fullest and possibly teleological realization of that interaction. Table 5.3 summarizes the argument's progression and indicates the focus of subsequent attention, where "+" indicates strongly present, "-" indicates largely absent, and "+ / -" indicates both presence and absence in significant but not necessarily comparable proportions within the practices of that culture. The inescapable elements themselves are poorly understood and require theoretical understanding, which to date partially tends to focus on the variable practices. Studying the inescapable, in its fullest realization, also may reveal the particular forms of value derived from the variable elements analytically excluded.

The significance of the particular inescapable elements identified can be strongly supported by analogies to orality emerging into literacy and a contrast with written literacy, or premodernity, as a dual triangulation effect.

Under orality emerging into literacy, selection labor concentrated primarily in memorization and communication of a linear spoken utterance, and selection power emerged as dialogue between human speakers. The primary analogy of orality emerging into literacy with modernity is that the entire linear utterance—now written rather than spoken—and the human labor embodied in it is potentially available for exploitation. Although the real possibilities of direct comprehension are constrained by the extent of available utterances, a compensating addition of computational procedures for selection has developed. However, those procedures involve tearing the word or other segment of utterance from its place in a linear utterance. The absence from modernity of an immediately human respondent in dialogue can be partially compensated for by the dialogic

Table 5.3
Contingent and inescapable elements in modern information retrieval

Practices				**Library and information practices**	**Internet practices**	
Contingent	**Variable**	Description labor	Human semantic description labor	+ / -	+ / -	
			Semantic search labor as description	-	**+**	
		Generic capacity offered by descriptions	Generic capacity from human description labor	+	+ / -	
			Generic capacity syntactically generated in description	-	**+**	
		Ordering of retrieval results	Intelligible orders	**+**	**+**	**Experimental information retrieval**
			Identical orders	-	-	
		Selection operations	Boolean operations for selection	+	+	-
Inescapable	**Common**	Selection operations	Computational procedures for selection	**+**	**+**	**+**
		Semantic	Production of meaning from written language	**+**	**+**	**+**
		Syntactic or pattern based	Frequency of occurrence of words and phrases	+	+	+

character of system interrogation, where queries can be revised on the basis of system responses. Selection labor—distinguished from selection processes—has begun to reemerge as a single category, although concentrated in searching rather than in description. The availability of the entire utterance for exploitation, the dialogic character of system interrogation, the concentration of selection labor as a single activity are then analogous between orality emerging into literacy and modernity.

Description products as records or metadata formed by human semantic labor on preexisting utterances were not part of information retrieval under orality emerging into literacy and can be regarded as practices developed under written literacy, with an often disguised connection to the constraints of the technologies of written literacy. Semantic human description labor and its products often are substantively absent or bypassed in searching under modernity. The human syntactic description labor of premodernity itself is transformed into machine description processes. The description products of modernity can then be regarded as partly specific to that era.

Therefore, substantive analogies that run across particular technological eras support the more methodological exclusion of human semantic description labor and its products and the focus on the linear utterance and procedures for selection from linear utterances in the subsequent consideration of full text retrieval (see table 5.4).

Table 5.4
Inheritances and transformations from orality and literacy

	Selection labor	**Linear utterance**	**Dialogic character**
Orality emerging into literacy	Concentrated in memorization and communication	Directly presented over time	Human respondent
Written literacy or premodernity	Development of description labor and products	Can be read	Absence of human respondent
Computational modes or modernity	Concentrated in searching	Available for exploitation, by computational (nonlinear) procedures	Dialogic character of system interrogation

Thus, the labor theoretic approach is both comprehensive and, to a certain extent, analytically revealing for full-text retrieval. Comprehensiveness was revealed by understanding that full-text retrieval is a product of the desire for selection power, enabled by modern information technologies. Human selection labor remains inescapable for producing selection power, and selection labor is concentrated increasingly in searching. Analytic power was revealed by precisely locating the primary dynamism for change and its secondary effects as a quantitative change in description processes that give rise to qualitative changes in human semantic search labor. Therefore, particular aspects of information retrieval identified by the labor theoretic approach have assumed particular importance (see chapter 4 conclusion). Within the identified locations for change, contingent and variable practices could be separated from inescapable elements, both in a synchronic analysis of current practices and from a diachronic perspective on historical inheritances. The labor theoretic approach was comprehensive in identifying the inescapability of selection labor and analytically revealing for isolating the focus for prospective attention, assisting the separation of inescapable from contingent elements.

For our purposes here, technological constraints on the quantitative scope of descriptions, often received as universals, have already been identified as transcended or no longer immediately relevant; we can analytically exclude them from the prospective discussion.

Prospect

We can maintain consistency with established theories for enveloping context as well as the specific approach to information retrieval, with a continuum between them. With regard to the broader context, the significance of the material basis of being to the making of human history, to consciousness, and to the possibilities of retrieval continues to be acknowledged. The material basis of being emerges specifically in the materiality of communication, particularly as a line of writing extended across a surface. Current transitions in information technologies are understood as a rematerialization and not a dematerialization of communication. With direct reference to the labor theoretic approach, selection power continues as the primary value and remains inescapably produced by selection labor, although human description labor is transformed extensively into a machine process and selection labor is concentrated

in searching. The labor theoretic approach identified the significant loci for change. Practical understanding remains valuable, and the articulated theory is deliberately congruent with the practical understanding of how to use systems effectively, particularly accounting for the perceived effectiveness and widespread use of phrase searching.

Congruent with a focus on the inescapable elements of information retrieval, some methodological restrictions are imposed. Human description labor and its products are deliberately and analytically excluded; similarly excluded are their analogs in generic capacity and the ordering of records or documents produced computationally by Internet search engines. Boolean operations for selection, highly common across systems and including the crucial factor of cutting a segment of written utterance from its line, are assumed. Methodological exclusion of the contingent enables focusing on the inescapable elements common to library and information and Internet cultures. At a further level of analysis, such exclusion could reveal the particular value added by contingent elements, thus explaining the transcultural analogs.

The selection of sources for understanding qualitative changes in the possibilities for semantic search labor is guided first by their promise of analytical insight. Existing sources acknowledged as relevant to understanding information retrieval, such as the work of Augustine and of Saussure, remain relevant and can be incorporated into the fuller understanding developed. In their current development, these resources have been exploited primarily for the negative antithesis to the nomenclaturism evidenced by Augustine and critiqued by Saussure, rather than a positive account of the production of meaning from written language. The danger of extreme relativism that might follow from a purely negative antithesis was emblematized in the fragile authority of Humpty Dumpty (box 5.1).

We require a fuller understanding that moves beyond predominantly antithetical contrasts to language as nomenclature and constructs a positive account of signification, or the production of meaning in written language. Understanding can be developed by a dialectic with existing sources recognized as relevant, particularly Saussure and Shannon. Aspects of their work can be synthesized, stressing materiality of communication and its current rematerialization and also stressing the significance of practical understanding already recognized in the labor theoretic

approach. Selection of elements of materiality from existing sources ensures theoretical consistency, which itself promises the further possibility of final simplicity and analytic value.

Conclusion

We used the labor theoretic approach to identify specific changes in description process and search labor located precisely within the categories offered by that approach; further development can be obtained by dialectical departure from sources already recognized as relevant. We still need to develop a fuller understanding of the inescapable elements of semantics and syntax common to both library and information and Internet cultures. The understanding can assist the theoretical integration that might provide a basis for informed practical action. We aim to develop a positive account of the production of meaning and the occurrence frequency of words and phrases, which can be exploited deliberately in searching and may also suggest teleological limitations on system development. While the account of meaning and of frequency will be congruent with the practical understanding revealed in known and widely diffused patterns of exploitation, it should yield fuller and more fully articulated understanding. Enhanced understanding should yield value for library and information and Internet discourse and practice. The limited sophistication of discussions detected here could be enhanced. In particular, theoretical constraints for future developments could be detected from the analogy of the syntagma with the message and the paradigm with the messages for selection. Linguistics could contribute an understanding of the role of the syntagma and paradigm in signification and information theory of the combinatorial possibilities of the messages for selection into the message.

We can now anticipate subsequent chapters concerned with full-text retrieval. Drawing on Saussurean linguistics, chapter 6 provides a materially embedded account of the production of meaning from written language. Chapter 7 discusses the recurrence frequency of words and of phrases. Chapter 8 synthesizes the insights obtained from both linguistics and information theory, particularly directly informed by an understanding of the computational process.

6

A Semantics for Retrieval from Full Text

Introduction

This chapter focuses primarily on semantics and draws specifically from Saussurean linguistics to describe the production of meaning in written language; it also provides a basis for understanding transformations of meaning in full-text retrieval. Meaning in written language will be understood as the product of human mental labor through interaction with the syntagma (the linear sequence of utterance) and the paradigm (the network of associations a word acquires outside its context in discourse). Syntagma and paradigm are fundamental to Saussurean linguistics and congruent with concepts from information theory: message and message for selection. We will use the concepts from information theory in the next chapter to account for the frequency of occurrence and recurrence of units and combinations of units in written language, principally words and multiword sequences. In addition, syntagma and paradigm are congruent with the material basis for algorithmic operations on the line of writing for description, searching, and retrieval. The congruence of fundamental categories implies that an account of the production of meaning, derived from Saussurean linguistics, may be particularly revealing for understanding full-text retrieval. While the account of the semantics of written language presented here is consistent with Saussurean linguistics, it must be extrapolated—not simply received.

Saussurean linguistics has influenced information science primarily through its incorporation into semiotics, briefly anticipated by Saussure (1916/1983, 15–17; Warner 1994, 9). Adopted by semioticians Saussure's distinctions regarding the syntagma as conceived in binary contrast to

the paradigm, and the signifier, sign, and signified are valuable for understanding and analyzing signification in written language. Further developed here, Saussure's distinctions yield an understanding of the effects of automatic indexing operations for full-text retrieval on the language of discourse, where words are characteristically wrenched from their syntagmas, effectively released into a paradigm (a network of associations)—where they can acquire a multiplicity of signifieds—and restored to a variety of syntagmas in retrieval.

Within information science, information retrieval research has given some attention to linguistics, but it has focused more on its mathematical and formalized aspects (represented by Zipf's work on the distribution of word forms in linguistic corpora and by its interest in syntactic structures and computational transformations which has some analogies to Chomsky's linguistics) (Zipf 1936; Montgomery 1972; Sparck Jones and Kay 1973). The resistance of Saussurean linguistics to mathematical or logical formalism (Harris 1987, xiv) may account for its limited adoption within information retrieval research. Recent data obtained from citation searching also indicate that Chomsky is more prominent than Saussure in English-language scholarship, with very limited intersection between citations to Saussure and to Chomsky (see table 6.1).

The primary motivation for information science's recourse to linguistics has been its search for effective transformational procedures on written language to aid automatic indexing and translation. Here, we aim to develop understandings that will assist direct human intervention; those understandings also may indicate the possibilities and limitations of formalization.

Table 6.1
Intersection of interests in Saussure and Chomsky

Citation made in 2004 in the *Social Science Citation Index* **and in the** *Arts & Humanities Citation Index*	
Saussure (all works)	64
Chomsky (linguistics not politics)	500
Saussure and Chomsky	6

Syntagma and Paradigm

For Saussure, the interaction between syntagma and paradigm—associative relations—was crucial to understanding language. Syntagmatic and paradigmatic relations constituted linguistic structure and determined how the language functioned (Saussure 1916/1983, 126). In contrast to the "interpenetration of morphology, syntax and lexicology":

> Only the distinction earlier drawn between syntagmatic relations and associative relations suggests a classification which is indispensable, and which fulfils the requirements for any grammatical systematisation. (Saussure 1916/1983, 135)/

In a development from the indispensability of the distinction, Saussure insisted:

> Everything in a given linguistic state should be explicable by reference to a theory of syntagmas and a theory of associations. (Saussure 1916/1983, 135).

From the perspective of modernity influenced by computational technologies, transformations on language involved in retrieval should be equally explicable by the reference to the syntagma and paradigm. We have already indicated the possibility of significant explication, understanding that creation of a full-text index involves tearing a word from its syntagma and releasing it into the paradigm. We also understand retrieval as reentering a variety of syntagmas.

In his *Course in General Linguistics,* Saussure is more concerned with paradigm and language than syntagma and utterance (1916/1983; Harris 1987, 125). Prioritizing language and paradigm implies objectification of human activity. In earlier neglected and difficult work on anagrams, the syntagma and its patterns of variation—conceived at various levels of granularity—received preferential attention (Starobinski 1979). In this context, the usual order of mention and consideration—paradigm before syntagma—is reversed, thus preserving the probable order in human history of activity and perception. The individual utterance is given priority over language as a whole (Vološinov 1929/1986), and the syntagma is regarded as an abstraction from its prior practical instantiation in oral speech occurring over time and written language extending across space. We can regard the paradigm, particularly as a network of associations, as a further abstraction, produced by the variety of syntagmatic occurrences of words. Finally, we will investigate the analytic value of the interaction

between syntagma and paradigm in understanding signification in written language. The discussion here accepts a degree of objectification of language but acknowledges and refers to the congealing of language in its written form. We further assume the removal of communication from direct semantic ratification—separation of the utterance from its place and situation of production and also from the possibility of questioning its producer.

Syntagma

Saussure considered linearity an inescapable and fundamental aspect of language, crucial to the conception of the syntagma. Linearity followed from the spoken nature of language, and the "spoken word alone constitute[d]" the object of study of linguistics (1916/1983, 24–25).

> The linguistic signal, being auditory in nature, has a temporal aspect, and hence certain temporal characteristics: (a) *it occupies a certain temporal space*, and (b) *this space is measured in just one dimension*: it is a line. (1916/1983, 69–70)

The principle of linearity is comparably significant as the "first law" of linguistics, the arbitrary nature of the linguistic sign (Saussure 1916/1983, 68–70).

> The whole mechanism of linguistic structure depends upon it [linearity]. Unlike visual signals (e.g. ships' flags) which can exploit more than one dimension simultaneously, auditory signals have available to them only the linearity of time. The elements of such signals are presented one after another: they form a chain. (Saussure 1916/1983, 70)

Therefore, linearity is fundamental to the concept of the syntagma.

> Words as used in discourse, strung together one after another, enter into relations based on the linear character of languages. Linearity precludes the possibility of uttering two words simultaneously. They must be arranged consecutively in spoken sequence. Combinations based on sequentiality may be called *syntagmas*. (Saussure 1916/1983, 121)

Some traces of objectification, particularly in the view of words existing prior to their instantiation in discourse, can be discerned in this passage.

For Saussure, the "spoken word alone constitute[d]" the object of study of linguistics (1916/1983, 24–25), although perception and understanding of the spoken word has been influenced by models in written language, both consciously and unconsciously (Harris 1987, 78). Both explicit and implicit models of written language appear in Saussure's treatment of lin-

earity, an inescapable and fundamental aspect of language inherent in the production of syntagmatic sequences. Saussure's idea that the linearity of the sounds of speech is most clearly evident when speech is transcribed to writing is highly revealing, exposing the hidden premise "of Saussurean linguistics . . . that the spoken word is 'invisibly' organized on exactly the same lines as the 'visible' organization of the written word" (Harris 1987, 77–78). Implicit influences from written language about the understanding of linearity can be discovered in metaphors of spatial extension—"a line, a continuous ribbon of sound" (Saussure 1916/1983, 102)—for the temporal linearity of the spoken signal. The auditory signal is conceived implicitly as abstracted from its producer and its visible supports and partly objectified.

The medium of discourse in which the syntagma is realized has further, more materially grounded, and empirically detectable effects. Biomechanics limit the combination of distinguishable sounds in the continuous or analog medium of oral speech. Saussure noted, "freedom to link sound types in succession is limited by the possibility of combining the right articulatory movements," and suggested, "[t]o account for what happens in these combinations, we need a science which treats combinations rather like algebraic equations" (Saussure 1916/1983, 51).[1] By contrast in written language, with its more discrete alphabet of symbols and potentially digital rather than analog nature, restrictions upon transitions between symbols are not inherent in the medium. However, restrictions can be imposed for certain communicative purposes, such as providing an analog to oral speech or incorporating redundancy to enable the reconstruction of messages disturbed by noise (Warner 2003).

Other analogies and contrasts can be made between the realization of the syntagma in oral and written discourse. The syntagma is extended in time in oral speech and in space in written language, although this contrast can be qualified. Perception—including reading—of written language may take place over time. Written utterances, such as private and published correspondence and issues of a journal, may take place successively and be related syntagmatically to each other. A further transformation of the relation between space and time occurs with the written text of a computer program: the program spreads out in text space and the computing process occurs over time (MacKenzie 2001, 38; Dijkstra 1968).

The linearity of speech becomes more evident when transformed into written form:

> This feature appears immediately when they are represented in writing, and a spatial line of graphic signs is substituted for a succession of sounds in time. (Saussure 1916/1983, 70)

The treatment of linearity includes explicit analogies with written language and implicit influences from writing. The particular conception of time as a sequence, with strong analogies to extension in space, is only one conception for time (Harris 1987, 77), which could itself be reinforced and made to appear natural by the diffusion of graphic representations of temporal sequences founded on the analogy of time with spatial linearity and directionality. Crucially for the purposes here, written language can be understood as a line sequentially composed of letters and other marks, including punctuation marks, and grouped into words divided from one another by spaces.

The linear materiality of writing is acutely revealed by an established pedagogic technique for exemplifying the distinction between syntagma and paradigm. Words are cut from a sheet of written paper,[2] a process more complex than it might first appear. First, the syntagma can be isolated by cutting the line of writing as a single, physically continuous ribbon from paper with conventionally arranged writing on one side or surface of the sheet of paper. Forming a continuous material ribbon requires a zigzag cut. In modern practice, the semiotic sequence of the line of writing is already broken at the end of each line, and these breaks are retained in the continuous ribbon. Historically earlier forms of writing—such as the boustrophedon, or the way the oxdrawn plow moves—would have yielded a simpler line of writing for cutting, semiotically continuous and more readily materially isolated. Once the syntagma has been cut as a continuous material ribbon, units of the paradigm can then be isolated and removed from it by cutting words from the line of writing. In the modern line of writing, words characteristically have boundaries, indicated by spaces or punctuation marks. Earlier forms of writing, including the boustrophedon, did not necessarily separate words and appeared as an unbroken line of writing more strongly analogous to the continuity of oral speech extended over time. The technique is rooted in the materiality of premodernity: the writing on the paper is the object of labor and scis-

sors are the instrument of labor. The operations of cutting can be based upon visually detectable patterns alone, without reference to meaning or semantics. Cutting can be accomplished computationally and assimilated to modernity by replacing human labor with a syntactic machine process that enables greater speed and a much more extended object of labor as the line of writing.

For our purposes here, explicit concentration on the line of writing replaces the analogy between the extension of oral speech in time and writing in space, with its partly covert premises. The line of writing is understood as materially embodied and extended across a surface, giving the possibility of moving backward as well as forward along the line. In language fully detached from its producer, the line of writing provides the closest practical realization of the form from which the concept of the syntagma can be abstracted.

The constituents of the understanding of the syntagma, word, and discourse demand consideration, including contrasts between written and spoken discourse.

Word

The word is an essential constituent of the syntagma; its definition has been disputed repeatedly in linguistics. For Saussure, "what a word is usually taken to be does not correspond to our notion of a concrete unit" (1916/1983, 103–104):

> To convince oneself of this, it suffices to consider the singular form *cheval* ("horse") and its plural *chevaux* ("horses"). It is commonly said that these are two forms of the same word. But, taking each as a whole, it is clear that we are dealing with two quite distinct items, as regards both meaning and sound.

Understanding of the word can be sharpened by considering this passage. As sounds and, more clearly evidenced, as sequences of letters, there is both resemblance (the letters, *cheva*) and differentiation (*l*, *ux*), but not the identity of a concrete unit between the two words. For meaning, or semantics, there is contrast between singleness and plurality, but also commonality of the constituents of reference (the horse as one horse and as one of more than one horse). Saussure continues to acknowledge the absence of "immediately perceptible entities" in language, but concludes by regarding the word as "a unit which compels recognition by the mind"

that "has a central role in the linguistic mechanism" (1916/1983, 105 and 109). Saussure recognizes but does not fully formally define words as units of written utterance.

From a materialist perspective, the word of written language could be regarded, in part, as a historically developed division of the written syntagma. Divisions between words would be understood as historically introduced between units that compel recognition by the mind. In this understanding, the word would be a possible origin for the perception of the paradigm, particularly for the paradigm as the network of associations created by the occurrence of identical or related word forms in a variety of syntagmas, with differing meanings or signifieds.

Beyond the Word

Saussure understood the notion of the syntagma as applying "not only to words, but to groups of words, and to complex units of every size and kind (compound words, derivative forms, phrases, sentences)" (1916/1983, 122). Units larger than single words—compounds and flexional forms (*il a été*, or he has been) (1916/1983, 104)—are then recognized. For Saussure, the word was an element of language and the combination of words in discourse would normally belong to speech. Like the word, some complex units (particularly compound words and derivatives) belonged to language—"to the language, and not to speech, must be attributed all types of syntagmas constructed on regular patterns" (1916/1983, 123). More extensive syntagmas, such as the phrase and the sentence, were parts of speech, although there was no clear boundary separating language or communal usage from speech, marked by the freedom of the individual. For Saussure, the "characteristic of speech is freedom of combination" (122), in a curious anticipation of the terminology used to describe the selection of messages from the source in information theory. Similarly, with "sentences . . . it is diversity which is predominant" (105). The syntagma then includes multiword sequences.

Paradigm

Saussure characteristically placed the paradigm (or, "associative relations") in a binary contrast with the syntagma (1916/1983, 123). The paradigm can be regarded as produced by and abstracted from the experience of the syntagma.

The paradigm has been understood classically in a triple sense:

- as the vertical axis, counterposed to the horizontal syntagma—a spatial rather than immediately temporal extension;
- as the collection of units or members of an associative group that can be substituted for one another in the syntagma, while remaining syntactically or semantically acceptable or cognate; and
- possibly most influentially, as the network of associations a word can acquire when considered apart from the syntagma, in accord with Saussure's primary understanding.

> Outside the context of discourse, words having something in common are associated together in the memory. (Saussure 1916/1983, 121)

The idea of memory as a network of associations entered modern cognitive science through the idea of the semantic network, without any discernible indebtedness to Saussure but possibly prompted by similar influences from language. The originally complex semantic networks generally have reduced to simpler and more computationally tractable tree structures that incorporate genus: species relations (Johnson-Laird 1988, 328–330).

Syntagmatic relations hold *in praesentia*; in contrast, associative relations obtain *in absentia* when the line of writing is considered alone. The syntagma immediately introduces the idea of a fixed sequence that contains a specific number of elements. An associative group lacks a fixed order and may have an indefinite number of elements. Some associative forms, such as the flexional paradigm of the cases of a noun or verb, may have a limited but not necessarily precisely agreed upon number of elements (Saussure 1916/1983, 122–124).

Representation of the paradigm on surface—as a diagram, rather than simply as line—is often essential for exposition; representations impose both pattern-based and semantic cohesion (Saussure 1916/1983, 125). Both the possibility of distributing cut pieces on a surface and the established form of representing the paradigm as a diagram on a plane imply that linearity alone is not sufficient to represent the paradigm in an immediately intelligible way.

The paradigm has been regarded as formed by units carved from the syntagma (Barthes 1984, 121). The metaphorical force of *carving out*

should be observed, as it may help understanding of the strength and extent of words released into multivalency when cut from the syntagma. Formation of the paradigm dividing the syntagma also has a historical analog in the introduction of word divisions into a previously undivided line of writing. A further common metaphor for the relation of the syntagma and paradigm involves weaving—the oral or written text as the woven product that is formed from the line of oral speech extended over time and of writing across space. In particular, written words removed from their syntagma can be regarded as torn from their line of writing and embedded in the fabric of the text.

From a rigorous materialist perspective, the paradigm can be regarded as generated from the syntagma, corresponding to the probable historical order of perception and the known historical order of the development of written language. In the pedagogic technique for exemplifying the distinction of syntagma from paradigm—cutting lines of writing from paper and isolating words from the line—the syntagma (the line of writing) can be cut repeatedly along its vertical axis, isolating words or elements of the paradigm. The possibility of cutting depends partly on word bounda-

Box 6.1
Tearing of a Word from Its Syntagm

> [T]he mandrake . . . gives a cry when it torn up; this cry can drive those who hear it mad. We read in Shakespeare (*Romeo and Juliet*, IV, iii):
>
> And shrieks like mandrake torn out of the earth,
> That living mortals, hearing them, run mad …

In German, "mandrake" is *Alraune*; earlier it was *Alruna*, a word that comes originally from "rune," which stood for "whisper" or "buzz." Hence (according to Skeat), it meant a "mystery . . . a writing, because written characters were regarded as a mystery known to the few." More simply, perhaps, the idea of a visible mark standing for a sound baffled the Nordic mind, and therein laid the mystery.

The physician Dioscorides (2nd century A. D.) identified the mandrake with the *circea*, or herb of Circe, of which we read in the tenth book of the *Odyssey*:

> At the root it was black, but its flower was like milk. Moly the gods call it, and it hard for mortal man to dig: but the gods are all-powerful. (Borges, 1974, 96-98).

The tearing of a word from its line of writing could be compared to tearing up a mandrake from the earth. Borges also revealed some direct association between the mandrake and the written word.

ries marked by spaces. By contrast, oral speech does not necessarily have intervals of silence (analogous to spaces in writing) between words, and it does not constitute a product that can be segmented, with the cut segments retaining material existence. Oral speech has no direct equivalent to the surface on which words can be distributed. Cutting the line of writing implies a disruption of linearity; the cut pieces can be gathered on a preexisting surface. Identical cut words from different syntagmas, possibly contrasting in meaning from their original syntagmas, can be regarded as the origin for the concept of the paradigm, particularly in the sense of a network of associations. Therefore, the paradigm can be regarded as produced by and abstracted from the experience of the syntagma.

The cut words can be distributed on a surface in ways that materially embody each of the triple senses of the paradigm. First, in the least deliberately imposed organization, cut pieces can be allowed to fall onto a surface as they are cut from above, similar to the material experience of cutting paper with scissors above a desk surface. Cutting along the vertical axis and allowing words to fall onto a surface corresponds to the first sense of paradigm as the vertical axis, counterposed to the horizontal syntagma, with a spatial rather than an immediately temporal extension in real material form. The technology of premodernity—paper and scissors—requires human material and syntactic labor. Under modernity, it could be conducted computationally, but the result would yield limited semantic interest.

With greater imposed organization, cut words could be grouped by similarity in patterns (for instance, by opening sequences of letters of words), resulting in pattern-based organization, where distinguished groups of words might constitute collections of grammatical variants (for instance, cheva*l*, cheva*ux)*. Such organization captures one aspect of the second sense of paradigm—the collection of units or members of an associative group that can be substituted for one another in the syntagma and still remain syntactically or semantically acceptable or cognate. This captures only that aspect of the associative sense of the paradigm that corresponds to similarity in patterns, not the semantic or even grammatical relations between elements dissimilar in pattern (consider *is*, *are*). Organization by pattern also captures the paradigm's material basis, considered as derived

from the recurrence of identical word forms in different systems, with contrasting meanings. Under premodernity, this form of organization could be imposed by human syntactic labor; under modernity, it can be generated computationally and is commonly implemented for retrieval from full text.

Finally, cut words can be grouped by referring to considerations of meaning, including similarity, opposition, and other connections. Grouping by meaning—semantic considerations—corresponds to an influential sense of paradigm as the network of associations acquired by a word considered apart from its syntagma, congruent with Saussure's understanding:

> Outside the context of discourse, words having something in common are associated together in the memory. (1916/1983, 121)

Grouping by reference to meaning is a form of semantic organization—drawing on human understanding and memory—and requires human semantic labor. The necessary presence of understanding and memory indicates a link with semantic networks but also reveals a contrast; the greater complexity possible in grouping cut pieces suggests that relations of meaning may be difficult to capture in computationally tractable models. Under premodernity, similar instruments and objects of labor (hand, scissors, and paper) could be used for semantic organization and were employed for syntactic organization. Under modernity, grouping by meaning cannot be implemented immediately by computation and would require continual direct human semantic labor, using a materially different instrument and object of labor. For forms of syntactic organization, the change in instrument (from scissors and hand to computational operations) and object of labor (from paper to electronic storage) enables greatly increased speed and scope.

Demonstrating the material embodiment of different senses of paradigm enables isolation of the specific sense of paradigm for retrieval from full text, without human semantic intervention in description. The primary sense is of the varied meanings acquired by a word in different syntagmas. This sense has a material basis, with contrasting instruments and objects of labor between premodernity and modernity. It embodies one aspect of the second sense of paradigm, referring to units or members of an associative group, specifically to an associative group of patterns

identical or similar to one another. It also corresponds to an aspect of the strongest sense of paradigm as the associations a word can acquire when considered outside a specific context in discourse or in a syntagma. In retrieval, the syntagma of search words are restored, and the diversity of meanings likely will exceed expectation. The searcher implicitly may have held a nomenclaturist model of meaning. In summary, search and retrieval reveals the distribution of specific units of the paradigm in systems; the human searcher can recover varied meanings, whose diversity can encounter a prior conception of meaning, possibly more restricted than intended.

We can also reveal the sense of paradigm uncaptured. Words connected in meaning but lacking similar patterns (part of the primary or fullest sense of paradigm) are not immediately recovered. Human semantic labor could be used to link such words together in either description or searching. This chapter has methodologically excluded human semantic description labor from the consideration of retrieval from full text, but it has exposed one highly significant source for its potential additional value, of linking semantically connected but syntactically unrelated units together.

Syntagma and Paradigm and the Production of Meaning in Written Language

Interaction between syntagma and paradigm can be made into a crucial element in a powerful semiotic model of signification. For Saussure:

> In its place in a syntagma, any unit acquires its value simply in opposition to what precedes, or to what follows, or to both. (1916/1983, 121)

Although *value* is a difficult and multivalent term in Saussurean linguistics (Harris, 1987, 118–123), it can be regarded in this passage as corresponding to meaning, or signified. The primary concern here is with the reception (rather than the production) of written language, separated from the possibility of immediate dialog with its producer. In receiving the written syntagma, the choice of attaching value to the word as signifier[3] would characteristically be guided by the signifieds or values attached to words preceding and following syntagma. Different languages would differ in the extent and position—before or after—of the relevant syntagma. Mark Twain suggested, "Whenever the literary German dives into a sentence,

that is the last you are going to see of him till he emerges on the other side of his Atlantic with his verb in his mouth" (1889/1996, 280). Saussure's understanding of the syntagma may have been influenced (partly unconsciously) by models in western and Romance languages.

Word Meaning

The model of signification that incorporates interaction between syntagma and paradigm can be applied to both the production of meaning from the individual written word and the sequence. Saussure both endorsed and developed the structuralist principle that distinguished units obtain their significance through their difference from proximate units—"*In the language itself, there are only differences*." (1916/1983, 118)—and applied this principle to understanding the meaning of words. Regarding the meaning of words as produced by mutual differentiations contrasts acutely with the ordinary view of language as a nomenclature. While opposing the related assumption in classic information-retrieval research that the word is a unit of meaning, it does correspond to the experience of searching full-text systems. In a dialectical development from Saussure, the surrounding syntagma guides the choice of signified from the paradigm that attaches to the word in written discourse.

Conceptions of language, implicitly but not necessarily explicitly accepted in ordinary discourse through their embodiment in the monolingual dictionary, can further illuminate the relation between word forms and their meanings. In a monolingual dictionary, words characteristically have many definitions attached to them, and the number of definitions is influenced by the depth and purpose of treatment. The relation was acutely noticed by the first modern English lexicographer:

> names . . . have often many ideas, but few ideas have many names. (Johnson 1755/1982, 15)

From the first part of this remark, "names . . . often have many ideas," the relation between a word form and its meanings can be regarded as a one to many relation. Therefore multivalency, or having many values, is confirmed by the monolingual dictionary as a condition of a word's meanings, with the meanings indicated by definitions given and fully displayed by illustrative quotations. The second part of the remark—"few ideas have many names"—implies that if ideas are extended to include

meanings, a meaning likely will be unique—recoverable from its illustrative display and open to elucidation, but not necessarily easily subject to decomposition—unless the meaning is deliberately compounded from simpler ideas.

On the model of singularity of meaning, synonymy—equivalence in signifieds between words—may occur only temporarily as a symptom of structural change. The observation that names have many ideas but ideas have few names was itself made in connection with questioning the existence of full synonymy:

> Words are seldom exactly synonimous; a new term was not introduced, but because the former was thought inadequate. (Johnson 1755/1982, 15)

Similarly, Saussure noted, "a language dislikes maintaining two signals for a single idea" (1916/1983, 162). Saussure often viewed language as an objectified autonomous force ("a language dislikes . . ."), in contrast to the earlier view of language as a product of human activity—"a new term was not introduced." The idea of absence of full synonymy is consistent with, or even a reasonable extension of, the idea that a word's meanings are produced by difference from other meanings in the language.

Therefore, word meaning is regarded as produced by human mental labor on the interaction of syntagma with paradigm. Words characteristically have many meanings and meanings may be singular, indicated by definitions but not necessarily easily further decomposed or reducible to simple equivalences.

Phrase

A crucial contrast emerges when considering the signification of syntagmas more extensive than the word, such as the phrase or the sentence. Similar forces of human mental labor acting on the interaction of syntagma and paradigm continue to apply, but their effects are different in the extended syntagma. Since the choice of signifieds is restrained by the further syntagma included, extending the syntagma beyond the word would be expected to reduce the multivalency of the constituent words. Indeterminacy— different interpretations—could remain. Again, full-text phrase searching would experientially confirm the reduction of multivalency and the persistence of indeterminacy.

Multivalency and Indeterminacy

The account of the mechanism for the production of meaning from written language can be made to yield a distinction between multivalency and indeterminacy. Multivalency (that is, having many values or meanings) is a condition of words used as ordinary written discourse. It is most acutely revealed when a word is torn from its syntagma— effectively released into paradigm—and reinserted into a number of syntagmas, from which we can legitimately infer different meanings. In a particular syntagma, the multivalency of a particular word is restrained by determining its meaning from the meanings attached to proximate words in the surrounding line of utterance. Indeterminacy, or difficulty in reaching definiteness in interpretation, may still exist as a remaining possibility. Once formulated, the distinction of multivalency from indeterminacy is simple and analyti-

Box 6.2
Indeterminacy and Multivalency

> Would you convey my compliments to the purist who reads your proofs and tell him or her that I write a sort of broken-down patois which is something like the way a Swiss waiter talks, and that when I split an infinitive, God damn it, I split it so it will stay split and when I interrupt the velvety smoothness of my more or less literate syntax with a few sudden words of bar-room vernacular, that is done with the eyes wide open and the mind relaxed and attentive.
>
> —Raymond Chandler. Letter to Edward Weeks, January 18, 1948 (Chandler, 1962/1997, p.77).
>
> The language used by Chandler is intended to give an appearance of precision, with restrained anger conveyed by the repetition of "split" and its assonance with the "it" of "God damn it." On a syntactic level, an appropriate oral reading would emphasize the separateness of words and their boundaries, or splits from one another, in the crucial passage. The meaning of crucial words, "velvety," "vernacular," and "attentive" is controlled by their syntagm—for instance, "*bar-room* vernacular"—and is not idiosyncratic. *Indeterminacy* has been deliberately avoided at the levels of expression and of meaning. Remaining indeterminacy could be regarded as a manifestation of the central paradox of writing, that its appearance of exactness is continually betrayed by the possibility of different interpretations (McKenzie, 1990).
>
> *Multivalency*, in distinction from indeterminacy, is still potentially present, with many values or meanings for crucial words possible in other syntagms. The potential for multivalency could be empirically confirmed by a search of full text resources.

cally powerful, particularly for understanding full-text retrieval, but seldom explicitly made. It could be expressed economically in terms of the interaction of more fundamental and materially rooted categories, without distorting those categories.

Conclusion

Therefore, we can state summarily the account of the production of meaning in human consciousness for the reception of written language already developed. In written language, word meaning is regarded as produced from human mental action—semantic labor—at the intersection of syntagma and paradigm. The experience of meaning, as distinguished from the mechanism for the production of meaning, is understood as an event in consciousness. The concept of meaning is not reduced to that of definition or paraphrase, and the meaning of a particular word is not identified with any acceptable definition or paraphrase given. Meaning is regarded as strongly analogous to the signified. The concept of the signified has the conceptual advantage of an established place in a series of systematic distinctions, particularly differentiated from signifier and sign, while the concept of meaning obtains greater richness from its resonance with ordinary discourse.

The conception of meaning developed is consistent with the existing sources recognized as relevant to understanding information retrieval, but also develops dialectically from them. The idea of language as nomenclature, with a one-to-one relation of word to object, continues to be rejected while the counter notion of meaning as produced by difference is accepted. In a dialectical development, the idea that language is not a nomenclature is no longer left as the partly unexplored term of a contrast or antithesis. The mechanism for selection from differentiating meanings is given a definite and materially specific form of human mental labor at the intersection of the syntagma with the paradigm. Possibly assisted by the focus on the material aspects of communication, the dangers of mysticism and extreme relativism, present in existing sources, have been avoided. The particular nature of the mechanism for selection from different meanings identified also has a significant analytic value from full-text retrieval, from its elements of congruence, with the current rematerializa-

tion of communication, in which computational operations of cutting are applied to the line of writing.

The conception of meaning developed here is consistent with the underlying presumptions of the labor theoretic approach but also translates them into more specific terms. Meaning as an event in consciousness is consistent with elucidation—rather than definition by decomposition—as the only way of giving verbal understandings of atomic facts or primitive terms. Meaning itself might be a primitive term. In accord with the underlying stress on the significance of the material basis of being, the mechanism that produces meaning—the intersection of syntagma with paradigm—has a specific material form.

In relation to classic information-retrieval research, relevance has not been identified with similarity in meaning or even seen as a simple function of meaning. The assumption of the value of delivering all (and possibly, *only* all) the relevant records is not resurrected. Rather, meaning is regarded as one highly significant focus or criterion for selection, possibly more fundamental than relevance. Considered apart from and after meaning, relevance might connect more directly with the ordering of documents or records in retrieval, analytically excluded from the discussion here.

7

A Syntactics for Retrieval from Full Text

Introduction

This chapter focuses on understanding written language at the syntactic level. The understanding of syntax adopted here refers to patterns, not directly to grammar, and is continuous and consistent with the understanding of syntax in the concepts of syntactic labor and machine processes, which were developed earlier as part of the labor theoretic approach. The continuity of the understanding of syntax implies the possibility that machine processes operate on the patterns isolated and the further possibility of revealing the basis for established computational operations, including those used in retrieval from full text. The specific concern is with the units of written language and their combination—the letter or character, the word, and the multiword sequence. We are further concerned with the frequency of recurrence of units and their combination, particularly of identical multiword sequences. The area of study from which the analysis will be derived involves information theory, rigorously interpreted and applied to written language. The stress on the materiality of communication and the primary focus on written language will be sustained. In contrast to Saussurean linguistics, information theory focuses exclusively on communication in its aspect as signals and not on its meaning. It can be adapted to yield particular insight into patterns of replication and difference in sequences of signals. Written language is understood as the message of information theory. Information theory offers a particularly revealing analysis of the units of written language and their combination, but it is seldom exploited for understanding the structure of written language in relation to retrieval.

The model of communication in information theory emerged from a mathematical culture that valued the precision and the possibility of mathematical development, obtainable from the abstraction of fundamental concepts and entities from everyday reality. The fundamental components of the model were identified by Claude Shannon during the late 1930s, and the model itself was developed and substantiated through cryptography during the 1941–1945 war.[1] Including both the model of communication and its mathematical development, the theory was finally made fully public in 1948 in a two-part article, "A Mathematical Theory of Communication" (Shannon 1948/1993). At a biographical level, Shannon had been interested in coding systems, including Morse code, during his adolescence (Liversidge 1993, xxii). On a cultural level, coding systems for transforming written language, including Morse code and other telegraphic codes, had proliferated during the late nineteenth century, partially resulting from the need to transmit messages over the geopolitical space created by the western expansion of the United States and its increased links with Europe (Warner 1993). Thus, practical understanding of coding preceded theory.

Data indicate a very limited intersection between citations to Saussure and to Shannon and that very few documents captured by Google (2005) include both the phrase *mathematical theory of communication* and the word *syntagm* (see table 7.1). Data is only suggestive, rather than conclusive, but it strongly confirms that the conjunction of information theory from the article "A Mathematical Theory of Communication" with the concept of the syntagma, pursued here, is almost entirely unanticipated.

Including its strong influence on the early development of information science (Roberts 1976), information theory has been widely diffused but not necessarily well understood since its formulation by Shannon in 1948 (1948/1993). Particularly in the humanities and the social sciences but not within information theory (Verdú and McLaughlin, 2000), Warren Weaver's interpretation in a highly influential introduction to the monographic publication, *The Mathematical Theory of Communication*, has been received mostly as explication, with limited recourse to Shannon's own arguments (Shannon 1956/1993; Tidline 2004). Weaver's introduction includes the suggestion that information theory can provide a model for understanding human communication at the levels of meaning and

Table 7.1
Intersection of interests in Saussure and Shannon

Citation made in 2004 in the *Social Science Citation Index* and in the *Arts & Humanities Citation Index*	
Saussure (all works)	64
Shannon (all works)	122
Saussure and Shannon	1
References on Google (5 May 2005)	
Mathematical theory of communication	53,500
Syntagm	10,300
Mathematical theory of communication AND syntagm	14

the effects of messages on their recipients and that it need not be confined to signal transmission (Weaver 1949; Fiske 1990).

The theory embodied in the article "A mathematical theory of communication" is valued directly, rather than analogically, for its scope. Under certain specified conditions, it is considered applicable to a variety of systems previously considered as separate entities. Since 1948, and particularly since the late 1970s, the model became relevant to the design as well as to the understanding of telecommunications and data storage systems. It currently is regarded as still valid and as setting fundamental limits for information or signal transmission (Verdú and McLaughlin 2000). The model of communication in information theory has antecedents traceable to Aristotle (Sperber and Wilson 1986, 5–6), and it resembles Saussure's less deliberately abstracted speech circuit (Saussure 1916/1983, 11–13). In its original intention, however, it was restricted primarily to communication as signals and not concerned with the understanding of messages.

This chapter preserves the level of concern with signals, and this helps clarify connections and contrasts between information theory and linguistics. In semiotic terms, linguistics was concerned with signifier, sign, and signified, both as complexes and at the levels of signifier and of signified. The primary concern here involves the relation of signifier to signified and the influence of the interaction of syntagma and paradigm on the realiza-

tion of that relation in written language. Some attention to complexes of signifier, sign, and signified was given when considering speech as marked by freedom of combination, where a partial anticipation of the combinatorial concerns of information theory was detected. Although Saussurean linguistics focused explicitly on spoken language its concepts could be and were applied to the understanding of written sequences (in chapter 6). In contrast, information theory focuses exclusively on communication as signals or on the signifier, and does not consider the relation between signals and meaning (see figure 7.1) (Shannon 1948/1993, 5). In its original formulation, information theory was also limited to discrete rather than continuous signals, although it recognized that continuous signals, such as oral speech, can be transformed into discrete form (Shannon 1948/1993, 7–8, 50, 75).

This chapter examines the application of information theory to written language in its aspect as signals, and develops from a suggestion—made by Pierce (1980) in one of the few but not only less technical expositions of information theory (Cherry 1978; Shannon 1968/1993)—that, in applying information theory to written language, we may be

> pushing a little beyond the mechanical constraints of language and getting at the amount of choice that language affords. This idea suggests views concerning the use and function of language, but it does not establish them. (Pierce 1980, 124)

Information theory is directly, rather than analogically, applied: the line of written language would be one instance of the message of information theory and can be revealingly understood from an information theory perspective.

Analogies are established between fundamental concepts from linguistics and information theory (that is, between the paradigm and the messages for selection and the syntagma and the message). The analytic advantages and interest of the analogies established follow from the con-

	Expression	Signifier	Signal	} Information theory
Linguistics	Relation	Sign		
	Content	Signified		

Figure 7.1
Levels of analysis for linguistics and information theory.

vergence of previously unrelated and deeply contrasting traditions. "A confluence of fundamental quantities always is intriguing" (Berger 2000, 661) and may yield insights not obtainable from either discourse in isolation. The intention, both in the direct treatment of information theory and for the analogies with concepts from linguistics, remains an enhanced understanding of the condition within which human choice operates, rather than implementing deterministic computations, beyond revealing the basis for Boolean operations. The emphasis on human choice is consistent with this book's treatment of semantics, but contrasts with the dominant deterministic approach to the use of information theory in retrieval.

Messages for Selection and Message

In information theory, the message and messages for selection are fundamental and interrelated components of a coherent and comprehensive model of communication; communication is understood primarily as the transmission of signals across telecommunication channels. In the model of communication developed for information theory, an information source chooses from messages for selection and then combines them into a message. In its technical and deliberately intratheoretic sense, information is a measure of freedom of choice in selecting messages from the source. Measures of information can be derived directly from the message if the stability of the statistical characteristics of messages is present or assumed across the samples analyzed. The message is then passed to a transmitter, which operates on the message to produce a signal for transmission before sending it across a communication channel; noise in the communication channel is assumed. The communication channel is linked to a receiver, which operates on the signal to transform it into a message that can be passed to the destination. The information source, transmitter, receiver, and destination can be a combination of human and technology. For instance, the information source could be a printer that selects messages from a type font, or a person who composes a telegram and implicitly chooses from the lexicon of the language, in accord with the combinative constraints of syntactically acceptable telegraphic messages. For a telegraphic message, the transmitter would be a combination

of telegraph operator and equipment, transforming the message into a signal for transmission; the receiver undertakes an inverse process, passing the message received to the destination (Shannon 1948/1993) (figure 7.2). The examples of printing and of telegraphic communication indicate the empirical actuality of the message corresponding to an object central to the understanding of the syntagma—the sequence of written language.

The concepts of the messages for selection and of the message from information theory can be materially illustrated in a similar way to the distinction of syntagma from paradigm, if the process of cutting paper is considered as reversed in temporal sequence and the cut words are assembled into the line of writing. The cut words correspond to the messages for selection from a source, conceived in this illustration at the level of granularity of the word rather than the character. The assembled line of writing corresponds to the message, which then can be transformed into a signal for transmission. If the assembled message requires a deliberately intended specific meaning, the selection of messages must incorporate semantic considerations. If there were no concern to produce a particular meaning for the message, selection could be conducted on the basis of patterns alone. The reversal of temporal sequence implies a degree of ontological priority for messages for selection over the message, and, accordingly, we will treat them first.

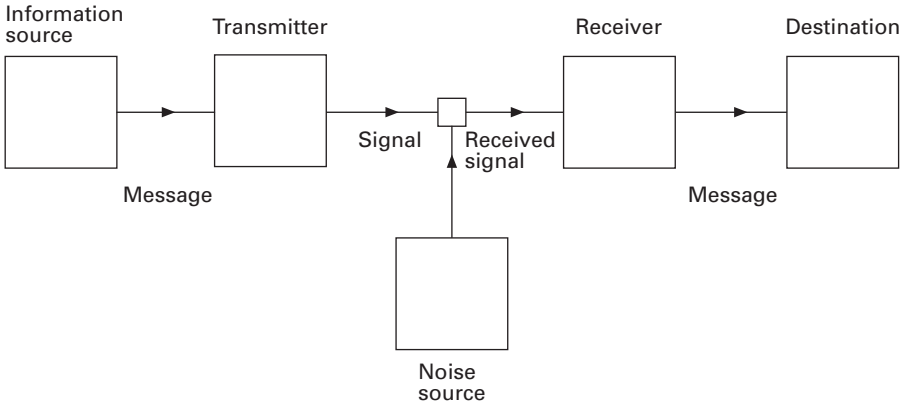


Figure 7.2
Model of communication in information theory.

Messages for Selection

In information theory, the messages for selection are the messages existing in the source from which the information source selects to compose a message. Classically, the messages for selection are conceived at the level of granularity of the individual character or letter of the alphabet of characters for selection (Shannon 1948/1993). Restricting the lowest level of granularity at which the messages for selection are conceived to the individual character avoids the necessity of direct attention to such features as intercharacter or mosaic differentiation and redundancy but still acknowledges the existence and function of these features (Cherry 1978). The level of granularity chosen is consistent with the understanding of the message adopted—orthographically acceptable written sequences from the English language lexicon, compounded of letters and other characters from the messages for selection.

Messages for selection can be conceived as held in a reservoir or scattered across a surface, in contrast to the surface organization often considered essential to representing the paradigm. Within information theory, the lack of organization implied in a reservoir or scattering across a surface is analogous to the formalized concept of entropy. Real historical analogues to a store, or reservoir, of messages for selection can be considered at increasing levels of organization within the messages for selection and not yet concatenated into a linear message, although possibly anticipating the relative distribution of the messages for selection in the message. Clarity in exposition can be obtained by restriction to a single alphabet (in this instance, the lower case Roman alphabet), although the concept of an alphabet could extend to other sets of characters. For instance, a relatively unorganized reservoir could be conceived as a container that holds equal numbers of individual letters. Combinations blindly chosen or randomly generated from the reservoir might be difficult to form into cohesive units or words in a given language, even on a purely pattern-based and computationally realizable basis, and a residue of unusable letters might remain after the formation of units (for instance, *qqq*, *xxx*, *zzz*). At a further level of organization, the distribution of letters in the reservoir could correspond to the relative distribution of letters in messages of the intended language; blind selections would then be more amenable to concatenation into lexically cohesive units or words, and there need only be a limited, possibly nil, residue of unusable characters. (*Scrabble*® players

should recognize a close correspondence with their modes of blind selection, combination into units, and ending with a limited residue). A still closed but partially organized reservoir could be constructed by organizing characters into categories, such as consonant and vowel, with blind selection from different categories possible: *Countdown* letter games could be understood as selection from categorized but effectively unlimited reservoirs (Countdown 2007). The collection of moveable type in a printer's shop could be regarded as an organized reservoir of letters open to deliberate selection of individual letters, with the number of letters in each box of the overall store anticipating their relative numerical distribution in message sequences (ligatures are excluded for the sake of analytical clarity but could be understood as frequently recurring sequences of individual letters). A surface with some pattern-based organization imposed would be the sending mechanism for the Cooke and Wheatstone telegraph, from which messages were selected by pointing needles at particular characters within an array. From a restricted alphabet of twenty letters, each character could be selected by pointing two of the five needles; characters arranged primarily by reference to the order of the alphabet (Ohlman 1996, 713–714; Science Museum 2007). A keyboard could also be regarded as a surface for selection—the historical transition from alphabetic arrangement to *qwerty* configuration was intended to ensure that type bars, linked to keys in a manual typewriter, did not conflict in their movements and jam (Ohlman 1996, 632). Proposals for reform from *qwerty* toward a more efficient keyboard could be understood as reorganization deliberately governed by frequency of occurrence of individual messages for selection in anticipated messages. In these instances, the messages for selection were conceived as syntactically or pattern-differentiated units, without immediate semantic significance. Individual messages only function semantically as words in special cases (*I*, *a*) and only when considered as part of the message.

The parallel between the increasing organization of messages for selection and of the components of the paradigm (Paradigm, chapter 6) is only partial. The organization imposed on the messages for selection is

- conceived primarily at the level of the character,
- restricted to pattern-based or syntactic characteristics, and
- stops short of semantic associations.

Semantic coherence could be imposed by direct human semantic intervention in the selection of specific characters from messages for selection for combination into a linear message, consisting of semantically cohesive sequences of messages as well as lexically acceptable units.

Messages for selection are analogous to the paradigm, particularly in its second sense of the collection of units or members of an associative group from which selection can be made to form a syntagma. Again, the paradigm has a semantic connotation when the messages for selection are syntactically conceived units. Reservations about the analogy arise in three respects:

- the assumption of preexisting messages for selection to the message,
- the level of granularity at which they are conceived, and
- the comprehensiveness of the messages for selection compared to the selective nature of the paradigm.

First, messages for selection are explicitly regarded as preexisting their combination in the message for transmission, in a further objectification of communication. In contrast, particularly for the interpretation here, the paradigm derived from the syntagma and existed partly *in absentia* and as an analytic construct. Therefore, information theory receives the products of human semiotic labor as objective existents. A real historical analogue to the independent existence of messages for selection can be found in a printer's font of separate metal types, where ligatures and the relative number of tokens in each category anticipated probable combinations and distributions in the message to be composed. Second, the messages for selection in information theory are conceived primarily at the level of granularity of the individual character; the character is regarded as a unit of an alphabet of characters that includes all characters admissible to the message. The constituents of the paradigm, understood as a collection of associated units, are primarily words—limited in number compared to the full vocabulary of the language. Finally, in contradistinction to the inclusiveness of the messages for selection, selection into the paradigm, based on associations in meaning or grammatical variations of a word or multiword sequence, is implied. Therefore, the analogy between the messages for selection and the paradigm must be qualified.

The concern here will be with the messages for selection, understood in an acceptable simplification as the individual and separate characters

of the Roman alphabet; the message is understood as syntactically and orthographically acceptable written sequences from the English language lexicon. The level of granularity at which the messages for selection are conceived stops at the individual character, without giving direct attention to such features as intercharacter or mosaic differentiation and redundancy, although the existence and function of these features is recognized (Cherry 1978; Warner 2003, 551). Thus, the message in information theory is understood as constructed from individual messages for selection and combined into a linear sequence or message.

Message

In information theory, the message emerges in two forms: the message for sending, passed from the information source to the transmitter, and the message reconstructed by the receiver from the received signal and passed to the destination. Classically, information theory was concerned with ensuring a close correspondence between the message sent and the message received, accepting that a signal would be perturbed by noise in the communication channel. The message is identical with the signal for certain uses of writing—handwritten, typed, or printed forms used for communication over distance—and neither transmitter nor receiver operates significantly on the message or signal. Selection errors and noise (for instance, orthographic mistakes and printing failures) are still possible, and some suggest that redundancy in written language was deliberately introduced to enable reconstruction of the message intended to be sent (Warner 2003). For other historically subsequent uses of writing, such as telegraphy and e-mail, the transmitter and receiver operate on the message and received signal, and technology displaces direct human labor and intervention over time.

We are concerned here with the message, not the transformations between message and signal. There are two methods for assimilating this concern to the model of communication in information theory. The full model could be retained, an identity between the message sent, and the message received assumed, with the transmitter and decoder working without error and the signal not perturbed by noise. Alternatively, the transmitter and decoder could be eliminated from the model and the signal regarded as the message in transmission or as the message as sig-

nal. Conceptually, eliminating the transmitter and decoder is preferable because it reduces the number of entities and corresponds to real historical developments, in which practical understandings of coding precede theoretical articulation, functions, and entities distinguished in information theory—such as the message and signal—separate from a previous lack of differentiation (Warner 2003). For the purposes of the current discussion, the message is understood as a single entity and no differentiation is made between the message for sending, the signal, and the message as reconstructed by the receiver; the transmitter and decoder are eliminated from the model. Although significant, these conceptual considerations do not affect consideration of the message's characteristics (including statistical characteristics). We will focus on choosing from messages for selection to construct the message and the implications of this process for understanding the structure of the message

Particularly on the understanding adopted, the message is both analogous with and can be differentiated from the syntagma. Shannon's references to the message as a sequence are reminiscent of Saussure's insistence that sequentiality forms the syntagma (Shannon 1948/1993, 6). The terminological similarity suggests conceptual congruence. A partly implicit rather than fully explicit assumption, rather than a principle, of linearity is similarly present in the model of communication in information theory, with communication proceeding in time from messages for selection to the destination. Linearity is implicitly observed in telecommunication practices for oral messages, presumably independently of Saussure's insight but influenced by similar considerations. If it is to remain ordinarily intelligible, an oral message must be received by the human destination in the same sequence as the message for transmission; graphic messages (including written language) can be received by the destination in parts over time, as long as an adequate correspondence to the message for transmission is finally reconstructed. Later modifications of the model incorporated feedback loops that allowed modification of linearity. In contrast to the concern of linguistics with levels of signifier, sign, and signified, information theory focuses exclusively on the signifier, or expression. Even when linguistics anticipated the combinatorial perspective of information theory in its view of speech as marked by freedom of combination, it was concerned primarily with complexes of signifier, sign, and signified, not the

signifier alone. Therefore, both the message and the syntagma can refer to a common object—the sequence of written language—with an underlying assumption or principle of linearity, albeit to different aspects of that object (the message to expression alone and the syntagma incorporating expression and content and the relation between them).

Once messages for selection are accepted as individual characters of the Roman alphabet, the message of written English can be understood at interconnected levels of granularity, differentiated as character or letter, word, and multiple or multiword sequences. In some special cases, an individual character demarcated by spaces or a punctuation mark and space can function as a word (the space itself can be considered as a character or message for selection, corresponding to the real material experience of selecting messages from a collection of moveable type or a keyboard). Both the word and the multiword sequence can be more fully understood from the perspective of information theory, specifically by treating the line of writing as the message of information theory.

Message and Messages for Selection and Occurrence and Recurrence of Units

From the interaction between messages for selection and the message, we can construct an understanding of the word and multiword sequence that is consistent with and complementary to understandings in Saussurean linguistics, but more fully developed. The intersection of syntagma and paradigm was transformed into a crucial element for a model of the production of meaning for written language, embracing and differentiating the word and phrase. Analogously, the combination of the messages for selection into the message can yield a syntactic or pattern-based understanding of the word and multiword sequence. The distinction between the levels of analysis of semantics from syntax is maintained and parallel and complementary understandings are developed. The messages for selection continue to be understood as individual characters, and the line of writing as compounded from the messages for selection, with the word as a demarcated group of characters and the multiword sequence as a linear concatenation of such groups. An informal, rather than arithmetically expressed or statistically formulated, but still sufficient understanding

of the frequency of recurrence of words, and particularly of multiword sequences, can be made to follow from these understandings.

Word

Shannon's own brief but highly illuminating explorations of the structure of written language, wherein the line of writing is effectively understood as the message of information theory, yield a definition of the word as "a cohesive group of letters [of printed English] with strong internal statistical influences" (Shannon 1951/1993, 197–198). This conception of the word is consistent with the Saussurean understanding of the word as a unit that compels recognition by the mind, but is more historically and medium-specific and less tautological. Shannon's definition of the word is specific to the written and printed word, not the oral medium, although it may have some application to handwritten utterances regarded as composed of discrete characters rather than as a continuous or broken line. While Shannon's definition of the word has been largely neglected since its publication in 1951, beginning signs of its adoption in connection with the computational parsing of language have appeared,[2] and it is potentially highly significant. Most fundamentally, Shannon's definition is congruent with an emphasis on the material basis for communication and its concentration on humanly and potentially automatically detectable patterns.

Each element of the definition can also be understood in a specific material sense, strongly related to linearity. *Group* can be understood as spatially grouped, particularly within the line of writing, and separated from contiguous words within the line by the space; the space itself is recognized as a character rather than simply the absence of a character. *Cohesive* can be understood in both a related and further developed sense of *group*—related, as the cohesion or mutual stickiness implied by the letters grouping together between spaces; and further developed, as cohesion between units of the word, or letters, which exhibit transitions between units, acceptable within that written language. *Strong internal statistical influences* imply that particular letters co-occur with one another, including transition probabilities between two individual letters but extending beyond immediately contiguous transitions to longer sequences.

Therefore, the definition of the word meets the rigorous test for knowledge as "an ideal reproduction of the external world serviceable for cooperative action thereon" (Childe 1956, 54). For instance, specific cooperative actions could include isolation of words from written discourse for full-text indexing conducted either by human clerical labor or machine process. Clerical labor effectively instantiated an understanding of the word strongly analogous to Shannon's definition, for instance for producing biblical concordances or nineteenth-century indexes for newspapers (Palmer 1885) before its formal definition. Clerical and automatic computational processes have used a similarly implicit understanding of the word and also have recognized the significance of the space; however, such understandings seem to have developed largely independently of the relevant theory. Thus, practical understanding has occurred partly in advance of theoretical formulation and diffusion, but practice could be informed by the encompassing theory.

The definition of the word as "a cohesive group of letters with strong internal statistical influences" is consistent with the already adopted conception of the message for selection as the individual characters of the Roman alphabet, and of the message as compounded from messages for selection. It also incorporates qualities specific to the message: cohesiveness and strong internal statistical influences. The definition of *word* can also provide a basis for a consistent understanding of the multiword sequence.

Multiword Sequence

As revealed by testing human subjects, prediction possibilities offered by the written-language message connect the word to and differentiate it from the multiword sequence (Shannon 1951/1993; Moradi, Grsymala-Busse, and Roberts 1998). From prediction testing, Shannon concluded that statistical effects extending up to one hundred letters (and therefore beyond the separate word) reduced entropy, the amount of information conveyed, to the order of one bit per letter for ordinary literary English. Understood as the negative of entropy, redundancy was roughly 75 percent; subsequent studies have yielded similar but not identical estimates (Moradi, Grsymala-Busse, and Roberts 1998). The space was revealed as almost entirely redundant in sequences of one or more words (Shannon 1951/1993, 198). Real-world coding systems have displayed a practical

understanding of the relative but incomplete redundancy of the space. While early forms of alphabetic written language do not mark word boundaries, the absence of the space symbol from Morse code required the introduction of further telegraphic transmission codes to prevent confusion over letter boundaries and word-endings in the message received (Warner 1993, 310–312). The space is a recent introduction to spell checkers for word processing systems. The reduced text produced by reducing redundancy can be considered an encoded version of the original:

> Roughly speaking, ideal prediction collapses the probabilities of various symbols to a small group more than any other translating operation involving the same number of letters which is instantaneously reversible. (Shannon 1951/1993, 204)

Analogously, shorthand—diffused in the late nineteenth century and considered here as a mapping from full to reduced written sequences, even if primarily intended for the transcribing of oral speech—reduced redundancy in the full sequence and was reversible. Reduced redundancy might introduce errors in reversing, particularly if the shorthand message was transmitted to a destination different from the information source and not used for private recollection over time, informed by knowledge of the circumstances of production of the original utterance.[3]

Shannon's experiments did not fully distinguish syntactic, or pattern-based, predictions from the semantic predictions of human subjects, introducing (possibly partly unconsciously) issues of meaning that are foreign to information theory's primary concern with expression. Thus, we can distinguish syntactic from semantic prediction and focus attention on the syntactic, or pattern-based, level already partially embodied in text compression programs. The possibilities of prediction are greatest with high redundancy and low entropy for subsequent characters, when sequences immediately below the word are considered: for instance, in written English the sequence *strengt* can normally only be followed by *h*. These predictive possibilities would be consistent with Shannon's understanding of the word as "a cohesive group of letters with strong internal statistical influences" (Shannon 1951/1993, 197–198). Compared with prediction from a combination of syntactic and semantic considerations, prediction based on purely syntactic features may cover shorter sequences and accordingly yield a lower value for redundancy and a higher entropy for written English.

From the perspective of information theory, therefore, the multiword sequence of written language can be understood as a linear concatenation of units weakly correlated with one another; the units themselves—understood as words—are internally cohesive. This conception of the multiword sequence is consistent with other aspects of the application of information theory to written language, specifically incorporating Shannon's definition of the word and building on the implicit assumption of linearity. Consistency with the definition of the word ensures further consistency with the conception of the messages for selection which were understood as individual characters of the Roman alphabet. This understanding was incorporated into the definition of the word. Coupled with effective real-world reference and, in this instance, computational applicability, internal theoretical consistency provides some assurance of the validity of the definitions of the word and multiword sequence. The understanding of the multiword sequence of written language is also simple, economic, and apparently novel, revealing simplicity and economy in the brevity of the understanding given. This impression of novelty is supported by its logical dependence on Shannon's conception of the word, itself not widely adopted despite its potential value.

The conception of the multiword sequence of written language as a linear concatenation of units weakly correlated with one another has other conceptual advantages, particularly through its correlation with other independently knowable features of the development of written language and related and independent theoretical perspectives. In relation to the historical development of written language, the space can be regarded as historically introduced and persisting where transition possibilities between two adjacent letters would otherwise be extensive and the potential for prediction of the second of the two letters from the first to be low, particularly for purely syntactically based prediction. After a space, the initial character of a word likely will be difficult to predict syntactically and will have low redundancy (some forms of writing—for instance, Biblical Hebrew—have indicated the presence of vowels at the beginning of words, but not within them). From the related perspective of critiques of the analogical application of information theory to semantic issues, the greatest freedom of choice at a semantic level occurs in the selection of the next word, where constraints at the syntactic level are weakest, reinforcing the evidence for the inapplicability of information

theory concepts directly to semantic issues. The conception of the multiword sequence is both consistent with and appropriately differentiated from independently developed concepts in Saussurean linguistics. The weak correlations between units correspond to the Saussurean view of speech as marked by freedom of combination. In contrast to conceptions of the extended syntagma and the examples given of extended syntagmas, semantic cohesiveness is not necessarily implied by an understanding of the multiword sequence in terms derived from information theory, which is limited to the level of expression. (It could apply to automatically generated sequences for which semantic cohesiveness was not attempted.) The conception of the multiword sequence has analytic value for understanding retrieval from full text and instrumental value for constructing specific queries when semantic cohesiveness is directly humanly imposed. In this understanding, a phrase that distinctively characterizes the topic desired is highly likely to occur only within relevant documents and this can be exploited in searching.

Differentiating the multiword sequence from the word as a weakly correlated concatenation of units that are themselves internally cohesive can yield a crucial theoretical insight into the frequency of recurrence of identical multiword sequences. If the units (words) of the linear multiword sequence correlate only weakly with one another, it is theoretically possible to recognize identical multiword sequences as highly improbable unless one is copied from the other. This conception of the multiword sequence also has considerable analytic power for understanding retrieval from full text—it can yield a theoretical explanation of the experientially encountered infrequency of recurring multiword sequences, even in very large corpora. The relative infrequency of identical phrases or extended multiword sequences, even in large corpora, can then be understood as consistent with the weak correlations between the concatenated units.

Summary

We derived a precise, computationally implementable understanding of the word and multiword sequence, realized in computational practice, by considering written language as the message of information theory. We also established implications for the frequency of occurrence of identical words and multiword sequences.

Conclusion

The specific interest of the analogies established between concepts from previously largely unconnected areas for full-text retrieval—Saussurean linguistics and information theory—lies in their distinctive possible contributions. Saussurean linguistics can give a basis for understanding and modeling the semantic effects of interactions between syntagma and paradigm, and the transformations of the signified or meaning in indexing, searching, and retrieval. Information theory gave insight into syntax or patterns, including a computationally implementable definition of the word and multiword sequence.

For the labor theoretic approach, we provided a basis for understanding the effects of the current transformation of selection labor toward searching. Emphasis on the material basis of communication, fundamental to the approach, has been sustained. The next chapter will theorize more fully the partly experientially and inductively acquired understandings of patterns of retrieval (given for the opening examples in chapter 4). The value of theory resides in the additional understanding it offers.

8

Semantics and Syntactics for Retrieval from Full Text

Introduction

This chapter is concerned with bringing the models of semantics and syntactics for full-text retrieval, developed in the previous two chapters, into dialogue with examples of retrieval from full text. To facilitate understanding of the material presented, we will maintain a strong parallelism with the structure of the previous chapters at macro-, intermediate, and micro-levels. At the macro-level, the material basis for communication is prioritized in the sequence of discussion, fully revealing the parallels between retrieval from full text and the semantics and syntactics already developed. At an intermediate level, we will maintain the sequence of examples given in chapter 5 and the progress from considering separate words to searching for phrases, adding a fuller example. At a micro-level, each example is considered first in relation to the specific semantics developed, then in connection with syntactics, and finally regarding other relevant considerations. We carefully preserve the already developed differences between semantics and syntactics and exploit their parallelism.

First, we must establish the correspondence between the already developed semantics and syntactics and retrieval from full text, primarily through reference to material basis.

Full-Text Description, Semantics, and Syntactics

Recurrent experiences in retrieval from descriptions of the full-text documents created by algorithmic transformations on the language of the documents themselves—both for transformations of meaning and for fre-

quency of occurrence of words and phrases—can be understood from the semantics and syntactics already developed. The semantics derived from Saussure can contribute to a model for understanding transformations as effects on the meaning or signified of document and query terms. The syntactics developed from information theory can give an understanding of the relative frequency of documents recalled by word and phrase searching. The congruences emerging suggest a validation of each approach. Contributions are obtained from both linguistics and information theory, and each discipline makes a distinctive contribution.[1]

Full-Text Description

We can derive a suggestive indication, if not confirmation, of the correspondence between language of discourse and language of representation for full-text indexing of syntagma and paradigm as well as the message and messages for selection from a further congruence between pedagogic techniques. We illustrated the relation of syntagma to paradigm by cutting words from a sheet of written paper. We could also illustrate the distinction between messages for selection and the message by considering the process of cutting reversed in temporal sequence. Similarly, the effects of full-text indexing can be demonstrated by cutting words from a sheet of paper and creating an ordered index from the cut words.[2] Both the cutting and the creation of an alphabetically ordered index can be conducted syntactically, working solely on patterns. The simplicity and materiality of the common method of exposition compels recognition of the correspondence and indicates its significance.

Computer operations represented by the pedagogic technique substitute a machine process for direct human labor and speed the task, possibly also accomplishing it more exactly. Although humanly constructed technology enables computer operations (Warner 2004, 5–35), the gestalt of the computer (Rosenberg 1974)—in the sense of a mysterious object held within relatively enclosed communities—may have disguised the simplicity of primitive transformations. The diffusion of technologies has rendered that aspect of the gestalt less potent, although it may remain necessary to reconstruct particular operations from the behavior of the system or from discursive account because we lack access to the source code.

Automata studies—a significant theoretical aspect of computer science—can contribute an understanding of the primitive computational operations possible on written language, also discovered independently in information retrieval. Computational operations can be understood as the writing, erasure, and substitution of symbols (Turing 1937; Herken 1995). Analogously for information retrieval from written documents, operations can be reduced to sorting or partitioning, which can be generated from writing, erasure and substitution, and the substitution of one symbol for another (Buckland and Plaunt 1994). The Boolean operators AND, OR, and NOT, familiarly used in commercial and Internet information retrieval systems, correspond to the primitive logical connectives,[3] themselves independently developed from the primarily iconic (rather than notational) modes used for automata theory (Minsky 1967, x; Warner 2001, 47–72). Different systems express Boolean operators as different commands, and the algorithms used in the experimental tradition also derive from the primitive operations.

A Semantics for Retrieval from Full Text

Saussure's remark that "Everything in a given linguistic state should be explicable by reference to a theory of syntagmas and a theory of associations" (Saussure 1916/1983, 135) can be developed to cover the transformations of meaning that result from algorithmic operations on language, as well as the production of meaning; this possibility will be indicated here. The associations or paradigm is understood in the strict sense of the possibilities of occurrence in the syntagma. By algorithmically creating a searchable description of a written document for full-text indexing, units of the syntagma are detached from their particular syntagmatic occurrences and effectively released into the paradigm (the paradigm is conceived as an analytic construct and *in absentia*). The level of granularity into which the syntagma is cut and the units are detached can vary, but the word, as a formal or pattern-based actualization of the word, has become a dominant unit in practice and will be the minimal unit considered here. Detaching units from the syntagma would cover a number of technically differentiated approaches, including serial searching of written language.

Therefore, searching operates on the units detached from the syntagma, in effect on units belonging to a particular paradigm or network of asso-

ciations. The minimal units isolated in description can be combined in searching (for instance, as direct Boolean combinations or as phrase- or sequence-searching, itself understandable as an application of the Boolean AND). Retrieval (the instantiation of the search statement) reinserts detached units into the syntagma, a reverse transformation of a paradigm into a multitude of syntagmatic occurrences. The signified, implied by the fuller syntagma revealed by this reversal, may not correspond to the signified intended by the search statement. An incomplete mental representation of the networks of associations for search terms may have been held, possibly due to the tendency to receive words from language as terms from a nomenclature.

In summary of the process of description, units of expression are cut from their fuller syntagma or sequence, where they may have had a definite signified, and released into a paradigm, where they become multivalent and admit a number of signifieds. In searching, a mental representation of the intended signified for the signifier, possibly received as having a single meaning (or as a *univocal* complex of signifier and signified), may be held. Retrieval restores the syntagmatic occurrences of search terms, where the signifieds represented by a single signifier may begin to constitute and even exhaust the variety contained within the particular paradigm. Therefore, linguistics can contribute a sophisticated understanding of the interaction between signifier and signified, enforced by the movement from syntagma to paradigm in description and from paradigm to syntagma in searching and retrieval, for computational and direct human operations on written language in full-text representation and retrieval.[4]

A Syntactics for Retrieval from Full Text

The message and the messages for selection are as fundamental to information theory as the syntagma and paradigm are to Saussurean linguistics. If the analogy of the message with the syntagma—and of the messages for selection with the paradigm—is accepted, information theory can be used to develop a theoretical understanding of the relative number of records or documents retrieved by different forms of searching, particularly the contrast between searching using word and multiword sequences. A more sophisticated formal understanding of the word, corresponding to its practical implementation, was derived from information theory.

As currently developed, automata theory and experience of information retrieval systems can continue to contribute an understanding of the effects of different logical combinations on the number of records retrieved, particularly for searches using combinations of separate words.

Exemplification and Discussion

We can now reconsider the examples given previously, in chapter 5, for searching on separate words and on phrases.

Word

The difficulty of finding two uniquely co-occurring words in the documents indexed by Google (*Unquiet cooccurrences / Quotidian metamorphise / Happiest legumens*) can be understood from linguistics, information theory, and the experience accumulated by using information retrieval systems (Google 2004). Linguistics, particularly through the idea of the paradigm as a mentally held network of associations, can indicate the difficulty of excluding semantic considerations from the choice of terms for searching for unique co-occurrences, comparable to the difficulty of humanly formulating a random sequence[5]—the repressed signified can reemerge. The operational implementation of the word—broadly, a sequence of characters between spaces or space and punctuation mark—is consistent with the conception derived from information theory (that is, a cohesive group of letters with strong internal statistical influences). The rarity of combinations of two such cohesive units, linked by AND, resulting in a single document can be understood in terms of the wide distribution of the cohesive units across documents. In addition, the rarity of combinations of two cohesive units resulting in a single document can be understood as continuous with but still contrasting to the historical experience of using information retrieval systems with documents recalled by simple Boolean combinations, particularly with reducing the number of documents retrieved by combining sets with AND. In a more deliberately theoretical, rather than directly experiential, perspective, combining entities assumed to be independently distributed with AND is analogous to multiplying odds or fractions. Fuller index descriptions and possibly transformations of the full texts of documents are now available in historical (including recent historical) experience.

Operating on full-text representations, the effects of searching using individual words and Boolean combinations of separate words, *(university AND library AND finance) OR (RAE AND research),* can be similarly understood. From the model of meaning already established, the restricted syntagma of the separate word allows a multiplicity of signifieds to attach themselves to the signifiers when they are effectively released into their paradigm; the retrieved extended syntagmas exhibit the multivalency acquired. The ordinary discourse meaning of words used as search terms was not stretched in either instance, and their use in retrieved syntagmas was not highly indeterminate. Initially, the inquirer may have held a nomenclaturistic mental model of query terms as univocal complexes of signifier and signified. The number of tokens of word types retrieved testifies to the preexistence of words as cohesive and discrete units in the language for representation. Using the technical affordances of the time, the syntagma was restricted to the paragraph for the Boolean combination *university AND library AND finance*, but the search recalled the Queen's Honors List (syntactically marked as a single paragraph), indicating the difficulty of defining the paragraph. Informal or experientially acquired and weakly theoretical understandings of the difficulties posed by the multivalency of separate words, embodied in facilities for restricting multivalency by the provision of proximity operators, again preceded their theoretical articulation.

Phrase

The effects of phrase searching—for example, *direct semantic ratification, Jack Kennedy was a friend of mine, Research Assessment Exercise* (Google 2004)—can be conversely and still consistently understood. From a linguistic perspective, the extension of the syntagma in the search and in the retrieved string beyond the separate word reduces the potential multivalency of the individual words (or the signifieds) that legitimately can be attached to them as signifiers, in so far as they are accorded separate signifieds. In the examples given, the search phrase is a relatively univocal (rather than potentially multivalent) union of signifier and signified, although the possibility of indeterminacy remains. Phrases could even be regarded as approaching semantic and possibly historically specific types of tokens retrieved (*Jack Kennedy was a friend of mine*) rather

than remaining simply syntactic or pattern-based types. Even in large corpora, the relative infrequency of identical multiword sequences can be explained from the understanding of the sequence developed from information theory—a weakly correlated linear concatenation of individually cohesive units. It is possible to make an analogy with cryptanalytic procedures, where repeated multiword sequences may be recognized, although possibly with access to the process as well as to the product of communication and directly to the place of the sequence within the encompassing utterance.[6] The utterance also emerges as more significant than the separate word, matching its possible ontological priority. Informal or experientially developed understandings of the effects of phrase searching—both as process of search, as with Google.com, and description, as with Amazon.com (Google 2005; Amazon 2005)—for both relative stability of meanings and infrequency of occurrence have preceded their theoretical explanation.

Fuller example

A fuller example can advance the dialogue between practical understanding and the theoretical knowledge developed. In the following extended sequence, *stood for* occurs in two deliberately contrasting meanings or senses, which could be paraphrased as "represented" and "put up with": "*A college widow stood for something in those days. In fact, she stood for plenty*" (McLeod 1932). In its first occurrence, *stood for* demands selection of *represented*, in the sense implied by the classic semiotic definition of the sign as something standing for something (*aliquid stat pro aliquo*), and in its subsequent occurrence, *to put up with*. The different syntagmatic occurrences compel selection of contrasting senses from the paradigmatic possibilities. The sequence *stood for* is regarded as belonging to the language, not to speech; from the Saussurean understanding of the word as a unit that compels recognition by the mind, it could also be regarded as a word. From the combinatorial perspective of information theory, the sequence *stood for* can also be regarded as a word if the space is treated as a character (Shannon 1948/1993; 1951/1993).

In its original filmic mode, by oral delivery, the particular selection from the paradigm is guided by the surrounding syntagma, but the process of selection is difficult to formalize in logical or computational terms.

Investigation could recover the paradigm for *stood for*, although the number of senses may be indeterminate and open to different imposed meaningful orders (for instance, by historical development or frequency of occurrence of different senses distinguished). Different surrounding syntagmatic sequences for *stood for* could compel similar or dissimilar selections. Sequences might be difficult to divide into mutually exclusive groups, demanding different choices: for example, what substitutions for *college widow* and *something* or for *she* and *for plenty* could compel similar or dissimilar selections from the paradigm as an association of signifieds for *stood for*? Even a mechanistic simplification and objectification of the process of selection—where the horizontal axis of the syntagma determines choice from the vertical paradigm—reveals a complex and computationally intractable process (see table 8.1). Even with a restricted example, human understanding of language appears to exceed logical and computational form—and is not necessarily reducible to such form—although human understandings can still create logical products.

A paradigm for *stood for* could be deliberately constructed by drawing on mental representations and associations in individual memory, while recognizing that these representations are historically and socially constructed, by drawing on reference sources, such as dictionaries, and on the collections of syntagmas derived from sources covered by Internet search engines. A historical perspective on the possibility of reconstructing the paradigm from external sources would reveal continuities and contrasts: continuities in forms of signification and contrasts in the type and quantity of direct human labor needed to recover the paradigm. Current information technologies, including Internet search engines, offer greatly

Table 8.1
Choice of signifieds for the term *stood for*

Syntagma				
Lexicon_1	Lexicon_2	*stood for*	Lexicon_3	Lexicon_4
Drawn from $\text{Lexicon}_{1..n}$		Paradigm for selection of senses for stood for, $\text{sd}_{1..n}$	Drawn from $\text{Lexicon}_{1..n}$	

enhanced possibilities for recovering syntagmatic occurrences from the line of writing.

An incomplete flexional paradigm (*stood for–stand for–standing for*) can be reconstructed from mental representations guided by historically accumulated grammatical understandings. The infinitive form, *[to] stand for*, which becomes the entry or head term for dictionaries, has no necessary special priority in the paradigm. As Saussure noted with regard to nouns:

> As far as language-users are concerned, the nominative is not in any sense the "first" case in the declension: the forms may be thought of in any variety of orders, depending on circumstances. (1916/1983, 124–125)

The need to reconstruct the paradigm implies that an incomplete mental representation of it may have been held on the semantic (rather than a syntactic or grammatical) level, possibly due to thinking of language as a nomenclature, with a single or highly restricted number of senses belonging to a word. Reconstructing the paradigm from syntagmatic occurrences confirms the existence of the paradigm *in absentia* and the syntagma *in praesentia* (Saussure 1916/1983, 122).

Monolingual and historical dictionaries of the English language reveal the diversity of senses that can be gathered under the verb *to stand* as headword. Johnson's subsequently influential approach to dictionary-making involved being empirical and inclusive, but he also used scholastic distinctions to inform differentiations between senses and the organization of definitions (1755/1982). *Stand for* is distinguished as a separate subentry within the entry for *to stand*, but neither *to represent* nor *to put up with* are distinguished as senses (Johnson 1755/1996; see figure 8.1). *The Oxford English Dictionary* also distinguishes *stand for* as a subentry, differentiating a greater variety of senses, including to "represent by way of symbol or sign" (the earliest occurrence of this sense is given as 1612) and to "put up with" (traced to 1896) (Oxford 2005). The senses of *stand for* distinguished primarily by Johnson are assimilated to senses of *stand*, with similar definitions and including similar illustrative quotations. Retrieval of *stand for* and *stood for* in definitions and illustrative quotation for other headwords—a form of retrieval that would have involved considerable direct human labor for Johnson and Murray[7] or their assistants (Johnson 1755/1982; Murray 1978)—reveals a further diversity

***Stand for* as entry term**

To STAND *for*. To propose one's self a candidate.

How many *stand for* consulships?—— three; but 'tis thought of every one Coriolanus will carry it. *Shakespeare*.

If they were jealous that Coriolanus had a design on their liberties when he *stood for* the consulship, it was but just that they should give him a repulse. *Dennis*.

To STAND *for*. To maintain; to profess to support.

Those which *stood for* the presbytery thought their cause had more sympathy with the discipline of Scotland, than the hierarchy of England. *Bacon*.

Freedom we all *stand for*. Ben. Johnson.

***Stand for* and *stood for* as syntagms in definitions.**

To ANSWER.

To be equivalent to; to *stand for* something else.

A feast is made for laughter, and wine maketh merry: but money *answereth* all things. Bible *Eccl. x. 19*.

APPE'LLATIVE. n.s. [appelativum, Lat.]

Words and names are either common or proper. Common names are such as *stand for* universal ideas, or a whole rank of beings, whether general or special. These are called *appellatives*. *Watts's Logick*.

OXSTALL. n.s. [ox and stall.]

A *stand for oxen*.

STILLING. n.s. [from still.]

1. The act of stilling.

2. A *stand for* casks.

Figure 8.1
Reconstruction of paradigm for *stand for* and *stood for*. Source: Johnson 1755/1996.

of senses, including a sense strongly analogous to *to represent as a sign* for Johnson (under *appellative*), as well as semantically dissonant senses with orthographically compatible signifiers in the definition of nouns (for instance, *oxstall* and *stilling*) (Johnson 1775/1996; a verbally similar definition of *stilling* is given by *Oxford* 2005).

A collection of syntagmatic occurrences of the written forms *stand for* and *stood for* can be obtained by an Internet search. In comparison to historical lexicographic enterprises, the Internet reduces direct human labor and alters the principle of selection. Literary sources no longer are favored, although selection still is confined to written forms. Selection is

Stand for Children

Building a powerful, grassroots citizen voice to ensure all children have the opportunity to grow up healthy, educated, and safe.
www.stand.org/ - 21k - 13 Apr 2005 - Cached - Similar pages

We **Stand for** Peace & Justice

... I **stand for** democracy and autonomy. I don't think the US or any other country **...** I **stand for** internationalism. I oppose any nation spreading an ever **...**
www.zmag.org/wspj/index.cfm - 12k - 13 Apr 2005 - Cached - Similar pages

BBC NEWS | England | Gloucestershire | New Kingsholm **stand for ...**

Work on a new £8m **stand for** Gloucester Rugby Club will start in summer, ready for next year's season.
news.bbc.co.uk/1/hi/england/gloucestershire/4428393.stm - 33k - 13 Apr 2005 - Cached - Similar pages

Lenin and Trotsky - what they really **stood for**

... what they really **stood for**. By Alan Woods and Ted Grant. Buy Online! Buy online from Wellred Books! Contents. A Note from the Authors **...**
www.marxist.com/LeninAndTrotsky/ - 3k - Cached - Similar pages

Crosses **stood for** women as well as babies - The Observer - Viewpoint

Crosses **stood for** women as well as babies, , The Observer, a newspaper of University of Notre Dame.
www.ndsmcobserver.com/news/2004/10/13/Viewpoint/Crosses.Stood.For.Women.As.Well.As.Babies-751052.shtml - 44k - Cached - Similar pages

Figure 8.2
Collection of syntagmatic occurrences of *stand for* and *stood for*. Source: Google 2005.

stood for something

The struggle for the GJF, A People's Foundation
... Justice and the Greensboro Massacre by Marty Nathan, MD & Paul C. Bermanzohn, MD. In Greensboro, North Carolina, an anti-Klan march and educational **...**
www.gjf.org/Klansmen.html - 31k - Cached - Similar pages
Volume 36 - continued
... In our team journal, there is an anonymous quote: "Live your life so your children can tell their children you **stood for something** wonderful." **...**
www.globalvolunteers.org/LINK/jumpvol36.htm - 12k - Cached - Similar pages
Netflix Customer E-mail 1/11/05
... America used to be a country that **stood for something**, ya know? **...** used to be a time when America, and American companies **stood for something** good. **...**
www.manuelsweb.com/netflix/0105email.htm - 14k - Cached - Similar pages
U2 Vertigo tour
... agree with a lot of his positions but at least he **stood for something**. **...** I mean, how can you respect someone just because he **stood for something**? **...**
flyertalk.com/forum/showthread. php?goto=lastpost&t=420174 - 101k - Cached - Similar pages
Hugh Duke 2nd Devons remembered
... and which he used of a fellow-officer who fell before him; "He **stood for something** very precious to me—for an England of my dreams, made up of honest, **...**
www.warchronicle.com/50th_div/ soldierstory_wwii/duke.htm - 12k - Cached - Similar pages

stood for plenty

Horse Feathers (1932)
... In fact, she **stood for plenty**. Frank: There's nothing wrong between me and the college widow. Wagstaff: There isn't, huh? Then you're crazy to fool **...**
www.filmsite.org/hors.html - 23k - Cached - Similar pages
tsujigiri: 01/02/2005 - 01/08/2005
... In fact, she **stood for plenty**. Lindsey Graham questioning Alberto Gonzales is like a doberman questioning Groucho Marx. Graham relies heavily on weird **...**
tsujigiri.blogspot.com/ 2005_01_02_tsujigiri_archive.html - 47k - Cached - Similar pages
tsujigiri: 02/06/2005 - 02/12/2005
... A college widow stood for something in those days. In fact, she **stood for plenty**. "And they who is not us shall perish"; Hoax: IQ in Blue vs Red States **...**
tsujigiri.blogspot.com/ 2005_02_06_tsujigiri_archive.html - 25k - Cached - Similar pages
[More results from tsujigiri.blogspot.com]
Fort Weyr
... He did become a Candidate and **Stood for plenty** of clutches. He finally met his lifemate, bronze Gehenth, and graduated from weyrlinghood, **...**
www.darkfort.co.uk/fort.php?p=persona2/ display_char.php&cname=K_sin&user=guest - 10k - Cached - Similar pages
2004 Pow Wow Page
... Like last year, the "POW" in "POW-WOW" **stood for Plenty** of Water, but that didn't stop the Rangers and Commanders of the West Central Section from being **...**
www.njwcsrr.org/NJWCS%20Pow%20Wow%202004.htm - 18k - Cached - Similar pages
Student Doctor Network Forums - Physical Medicine & Rehabilitation **...**
... Answering "because I heard PM&R **stood for plenty** of money and relaxation" is probably not going to earn you many points. **...**
forums.studentdoctor.net/showthread.php?t=187469 - 89k - Cached - Similar pages

Figure 8.3
Collection of syntagmatic occurrences of *stood for something* and *stood for plenty*. Source: Google 2005.

understood to occur on the basis of congruence of characters between query and document strings, with a further algorithmically generated ordering of references based on such features as number of links and recency. Inclusion of the context for the query string in the display of references embodies a practical understanding of the syntagma's value in reducing multivalency. Instances of *stand for* and *stood for* are difficult to assign exclusively to the sense distinguished lexicographically, although a sense of *to represent legitimate interests* seems numerically dominant, and *stand for* is also present in its semiotic sense (see figures 8.2 and 8.3).[8] Indeterminacy can remain, and multivalency has been reduced. *Stand for* also occurs with *stand* as a noun. The display is indicative of the complexity that might confront a lexicographer. The witty clarity of the initial contrast—*stood for something … stood for plenty*— is confirmed as imposed rather than emerging.

If the extent of the syntagma for matching is increased to *stood for something* and *stood for plenty*, multivalency of items retrieved are further reduced, although *plenty* still occurs as use and mention (figure 8.4). The number of occurrences of each multiword sequence and of documents retrieved is also further reduced, explicable by an information theory perspective on the occurrence and recurrence of words and multiword sequences.

The wit of the utterance, written but intended for oral delivery, stems from the clarity of the contrast between the two senses and the rapid transition from one to the other. The association between the two senses may have its origins in the unconscious (Freud 1905/1976, 215–238), but their mutual contrast is deliberately intellective. The contrast between the word senses also embodies reservations about attempts to construct a semantic unity or even strong coherence for a word:

> The various contexts of usage for any one particular word are thought of as forming a series of circumscribed, self-contained utterances all pointed in the same direction. In actual fact, this is far from true . . . Contexts do not stand side by side in a row, as if unaware of one another, but are in a state of constant tension, or incessant interaction and conflict. (Vološinov 1929/1986, 80)

The wit arises from the tension between the two senses and is lost when their order is reversed. Experience with full-text retrieval has tended to offer empirical confirmation of the variety of unpredicted syntagmas for words.

Extremes of relation between expression and content can be identified, and a progression can be traced between them. At one extreme, the sequence cut can have a multitude of signifieds: for Vološinov, "*Multiplicity of meanings is the constitutive feature of* [the primitive] *word*" (1929/1986, 101). In this understanding, the primitive word was not distinguished from the utterance or abstracted from its place between speakers. For an individual speaker, brevity and multivalency can reduce labor in the production of utterances, without demanding compensating labor in interpretation by that speaker, particularly if there is a power relation over the hearer, such as the slave, or "*instrumentum vocale*" (Marx 1867/1976, 303n–304n). In a speech circuit, or with the circulation of linear written messages, speakers are also hearers and have an interest in reducing the work in disambiguating multivalency in interpretation, bringing it into balance with work in production. Multivalency can still remain, particularly when the word is extracted from the speech circuit or written message. At an antithetical extreme, a signified could have many signifiers, although this would demand unnecessary labor both from source and destination. Saussure strongly questioned the existence of full synonymy, a manifestation of a signified with more than one signifiers (1916/1983, 162). A nomenclature with a unitary relation of signifier to signified would occupy an intermediate position between two extremes.

The fuller example has revealed that human understanding and composition of written language may be highly intractable computationally, even given sharp differentiations between alternative senses and with comprehension and selection reduced to a rather objectified form. Derived from a monolingual dictionary, the paradigm for the infinitive form (*stand for*) may give a sense of organization, clarity, and strictly limited number, and also of words or messages for selection that await combination in the syntagma or message. The monolingual dictionary has been regarded as a practical anticipation of the Saussurean model of language; the headword and definition are analogous to signifier and signified, and the network image of the semantics of a language is exemplified by the mutual definition of words. The collections of syntagmas revealed the diversity of speech, which may have been disguised by the appearance of finished order given by the collection of paradigms, and they indirectly testified to the abstracting and ordering efforts of lexicographers and theorists. From

a Saussurean perspective, the sequence *stood for something* could be regarded as partly belonging to the language—from its regular pattern—and partly to speech, while *stood for plenty* is clearly part of speech. For information theory, the contrasts in the relative frequency of occurrence of the two different messages would be consistent with the strength of the semantic, rather than syntactic, associations between internally cohesive units.

Summary

Thus, our review and analysis of examples has indicated the value of understanding signification in written language and transformations on written language from the interaction between syntagma and paradigm. The understandings already developed imply limits to logical or computational formalization. The syntagma has been consistently present as a real existent, while the paradigm has tended to be realized as an analytic construct. Practical understanding of the value of the extended syntagma, both for searching documents and for displaying references, has thus far largely preceded any theoretical articulation.

Conclusion

We have derived specific analytic advantages from the congruences identified between linguistics and information theory. Linguistics has provided a model for describing the relation between signifier and signified, when the signifier—as word or multiword sequence—is extracted from the syntagma, released into its paradigm, and reinserted in a variety of syntagmas by computational and human operations on written language for information retrieval. The most distinctive and valuable contribution has been an enhanced understanding of the shifting relation between expression and content.

We developed an account of the production of meaning in written language as interaction between the syntagma and paradigm. We approached, but did not fully exhaust, the limits of choice offered by written language and its computability on a semantic level. The preference implied by common patterns of interactive use for nondeterministic modes of use of systems could be strongly supported. The convergence of Saussurean

linguistics and information theory, informed by an understanding of the primitive computational operations, suggests constraints on deterministic selection from written language relevant to information retrieval. At this stage of conceptual development, constraints can be stated only as provisional generalizations, not confirmed. To create index descriptions, deterministic operations on written language may involve tearing words or other syntactically differentiated units from the syntagma and releasing them into paradigmatic multivalency. Multivalency may emerge most clearly in syntagmas retrieved by Boolean operations in retrieval, but it could remain present with other algorithmic procedures for description and retrieval. Extending the syntagma cut in description or searching can reduce the potential for multivalency. A preference for nondeterministic modes in searching can be supported theoretically as well as experientially. The human capacity for choice—earlier connected with the etymology of *intelligence* (*inter-legere*, or to choose between) (Scholarly and Ordinary Discourses, chapter 2)—exceeded the possibility of computational modeling, particularly for combining and understanding the syntagma. Information theory yielded an understanding of the relatively frequent occurrence of word and multiword sequences, extracted from the message.

The respective domains of linguistics and information theory have been sustained, although some convergence emerged from the dynamics of each discourse: in the linguistic view of speech, as marked by freedom of combination, and in the elements of semantic selection by the source and interpretation by the destination in information theory; Shannon's own experiments on language did not separate semantic from syntactic processes (1951/1993). Insights into full-text retrieval have been obtained from the correspondence of the syntagma and paradigm, and the message and messages for selection, to processes of representation in full text retrieval, which could not have been obtained from any of the discourses in isolation. The computational transformations possible on written language continued to be understood from automata theory, with an independent and consistent account given by information retrieval experience and research. Practice and theory have been brought into a dialectical relation to each other.

9
Conclusion

Introduction

This chapter will review the semantics and syntactics already developed in relation to:

- preexisting theories relevant to the information retrieval (outlined in chapter 1),
- the labor theory approach, and
- existing and emerging real-world practices.

Since the labor theoretic approach and the understanding of semantics and syntactics are both relatively novel, the theories developed in this book may not have explicitly informed real-world developments. If the theories have validity, however, real-world systems inevitably are both constrained and enabled by considerations connected with human labor (its costs, the ability of human semantic labor directly to address the level of meaning, and the possibly of transferring syntactic labor to technology) and also by inherited modes of production of meaning from written language and the nature of the computational process. We will indicate a potential synthesis of library and information science with Internet cultures that emerge in practice before being articulated in theory. Finally, this chapter will reaffirm the value of selection power and the merging possibilities for enhanced selection.

Theories

The order of consideration for theories and practices preserves the order of presentation throughout the book, beginning with the still-dominant tradition of classic information-retrieval research.

Classic Information-Retrieval Research

The development of experimental information retrieval systems usually has occurred independent of linguistic theory. In the early 1970s a review concluded, "The most striking fact to emerge from the literature, however, is the difficulty of marrying linguistic techniques and retrieval objectives" (Sparck Jones and Kay 1973, 197). Linguistically, very crude procedures seemed to work quite well for retrieval (understood primarily as the transformation of a query into a set of records) and it was unclear what more sophisticated procedures could contribute (Sparck Jones and Kay 1973, 197). Similarly, indexing solutions given to selection problems from natural language owed very little to linguistics (Gardin 1973, 140). Prior to more widely diffused interest in metalanguages during the late 1980s, only library schools had accepted that "[humanly assigned] retrieval terminologies . . . [were] worthy of sustained study and research" (Roberts 1989, 103). An understanding of languages of description—particularly those produced by algorithmic transformations on the language of discourse—had to be constructed from linguistics rather than imported as an established product. In this context, this book has addressed the Janus-like character of information—familiarly regarded as facing both the technical world of bytes and data compression and the social world of language and meaning (Gregory 2005)—that also requires, equally significant but less fully addressed, understanding from the human and discursive as well as the mathematical and computational sciences

The value of classic experimental information retrieval research to current practice and understandings is further reduced by the understandings developed. We must question the historically inherited preoccupation with the word as a unit, and with the statistical distribution of word forms (Zipf 1936). The received idea of the word as unit of meaning should be abandoned finally and replaced with a more sophisticated model of signification. The degree of success of information retrieval based on the word unit, and the continuing utility of individual word-based techniques, results from the coincidence of the language's semantic component, even if understood as an associative paradigm with syntactically separated sequences of the message. The utility for retrieval of combining established techniques with developing understandings of the stability of meaning and the frequency of occurrence and applying them to more

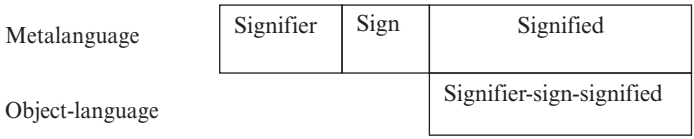
extended sequences from the syntagma and the message—to the phrase and the multiword sequence—has been demonstrated.

Less explicit assumptions of information retrieval research also demand reconsideration. The view that an index description must be briefer than the document described can be regarded as a historically specific product of the storage constraints imposed by print and early modern computer technologies. Rather neglected considerations of the depth of indexing could be revisited. Most strongly, the assumption of deterministic modes of interrogating systems, which correspond to the batch processing of the 1950s, can be replaced by a theoretically supported preference for nondeterminism. The preference for nondeterminism concurs with advocacy of selection power over query transformation as a design principle for information retrieval systems, particularly when selection power is conceived as property of human consciousness rather than of system processes.

Library and Information Science

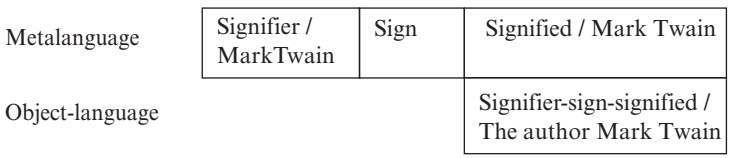
From the perspective of library and information science, exposing humanly assigned metadata as a historically specific development of premodernity—not fully present in orality emerging into literacy but persisting into modernity as a renewed inheritance—was crucially relevant for the labor theoretic approach and the practice of information retrieval. Such recognition enables theoretical separation of the current value of metadata from its persistence as an inheritance of modernity. In a synthesis that combines recognition of humanly assigned metadata as a specific historical phase with the semantics developed, such metadata could be regarded as gathering together different units from the syntagma, yielding generic power (or in some less frequently occurring applications, differentiating identical units of expression) and thus enhancing the possibilities of specificity in searching (figure 9.1). We suggested here that humanly assigned metadata could maintain its value for generic capacity (that is, collecting together disparate items or aspects of items from the original discourse). Currently, the separation of value from inheritance has occurred more frequently in practice than in theory, which has been constrained by intellectual inheritance and may require further transformation (Wilson 2001).

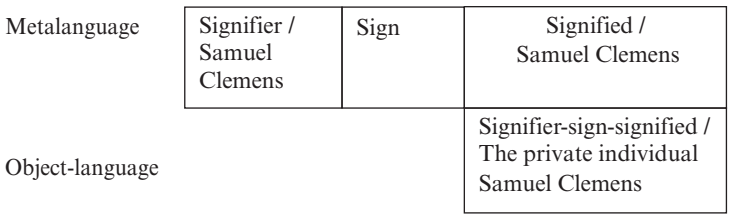
A metalanguage describing written documents taken as an object-language would treat the entire complex of signifier, sign, and signified as its signified.



This pattern might be known as a double articulation.

If the metalanguage were algorithmically generated from the object-language it could contain the phrases, Mark Twain and Samuel Clemens, and search could be highly specific (but could only link the two aspects of the single individual together by listing those terms).





If the metalanguage were purely humanly assigned (as would have been the case with paper-based indexing), it might not distinguish between Mark Twain and Samuel Clemens, but gather all occurrences of these terms under the heading, Mark Twain, giving generic power in searching (but losing specificity).

Figure 9.1
Meta- and object-language.

Consistent with the recognition of humanly assigned metadata as a specific historical phrase and corresponding to existing familiar (if informal) practice, a transformed relation between assigned categories and language of discourse could occur. In searching, chronology or temporal sequence could change from first using metadata and then considering the language of the original discourse to searching first on descriptions syntactically generated from the language of discourse and then exploiting the products of human semantic description labor found in association with the desired language of discourse, possibly for their generic power. An analogy exists with the established (but often informal) practice of identifying an item of interest and then looking at items categorized close to it—browsing enabled by collocation or human description labor. The temporal sequence of search corresponds to that of description, reinforcing the analogy between description and searching, and is analogous to the real historical process that moved from linear utterance to briefer descriptions of utterances. Giving attention first to the language of discourse— the full text of a work—exploits a resource characteristically produced by human semantic labor of greater duration and intensity than metadata, *prima facie* making it the richer resource. A crucial value for human semantic description labor—the gathering together of items related in meaning but dissimilar in pattern—is also exposed. Practice increasingly realizes this form of value.

Information Society

The concept of informational labor was adopted from information society discussions, with the intention of further differentiating the concept. A category of *mental* rather than *informational* labor was developed, and substantial conceptual advantages have accrued. The concept of *mental* labor implies continuity with what existed before modern *information* technologies, themselves produced by mental and physical labor. A sharper distinction between human and machine processing is implied, contrasting *mental* (specifically human) with *information* (human and machine). Within human mental labor, valuable distinctions between semantic and syntactic labor could result from coupling the concept of mental labor with established and historically warranted distinctions between levels of analysis. The differentiation of semantic from syntactic labor contains

a distinction between human labor and potentially machine processes: human syntactic labor potentially is transferable to technology, where it becomes a machine process. These conceptual advantages imply that the concept of mental labor is preferable to informational labor. The value of the further distinction made between semantic and syntactic mental labor was revealed by its ability to meaningfully comprehend real world practice.

Labor Theoretic Approach

The labor theoretic approach remained analytically revealing and sufficient to comprehend retrieval from full text, even when the activity of human description labor (which prompted development of the approach) was analytically excluded. Comprehension was revealed by the fundamental proposition that selection power is produced by selection labor, which becomes concentrated in searching as human semantic labor. Distinct from labor, syntactic description and search processes could be transferred to technology. Analytically, gathering together items related in meaning but not strongly connected in pattern was revealed as a unique function of semantic labor not easily (if at all) obtainable syntactically, which could have value in enhancing selection power. Attention to the costs of human labor in the labor theoretic approach reassumes relevance if the analytical exclusion of human description labor is removed, accounting for the distribution of the products of human description labor and the transference of labor to searching, and thus implying the continuing development of these practices. The semantics developed for retrieval from full text also may be relevant to the labeling decisions made by a describer concerned with meaning of a document.

While continuing to concede that changes in practice precede theoretical explanation and that practice may embody deep forces and constraints that are acknowledged but not explained away by theory, the value of the theories developed can be isolated and recognized. Theoretically understanding a dynamic offers the possibility of informed intervention in practice, particularly in the areas of freedom left by the determining forces. Transference of syntactic processes to machine may be inevitable, and it yields benefits for selection power. Since the distinctive value of human semantic search labor in description and searching—giving generic power—has been isolated, it can be deliberately exploited.

Internet Cultures and Practice

Practical systems encompassing objects without humanly semantically assigned metadata or without direct and independent access to that metadata (where it exists) have proliferated. Systems inherit patterns of signification from orality as well as from written literacy, suggesting historically deeper and socially wider roots compared to humanly assigned metadata. Patterns of signification commonly encounter with computational processes and a crucial tension exists between the linearity of writing and cutting from a line. Through both direct use and the secondary effects of their use, practical systems mediate to effects outside information retrieval.

Practical understanding of appropriate and market-accepted techniques for information retrieval from written language has preceded and seemingly is independent from the theoretical elements derived here from Saussurean linguistics and information theory. However, analogies between practices and theory and the possibility of more deliberately and theoretically informed developments could be established. The assumption that relevance is a function of the meaning of verbal forms can be qualified by recognizing the influence of such features as quality, with consensual notions of quality implemented in page-ranking techniques and the significance of socially distributed knowledge embodied in such techniques anticipated by social epistemology (Shera 1952/1965). Google's practice of displaying the surrounding line of search terms in retrieved references embodies an understanding of the value of the immediate syntagma in initial (but not necessarily final) disambiguation (Google 2005). The address of documents might indicate the place of utterance. Indicating the place of utterance is congruent with (but in the absence of stated addressees does not completely fulfill) Vološinov's, rather than Saussure's, insistence that a word obtains its meaning from its position in a dialogue between speakers (Vološinov 1929/1986, 65–123). The possibility of exact phrase searching on Google.com and the use of statistically improbable phrases by Amazon.com corresponds to theoretical recognition of the distinctiveness of content and the relative infrequency of occurrence of specific sequences of more than one word. Searching in Google can exploit these features; for Amazon.com, the features are presented as products of computationally generated descriptions that demand contrasting distributions and types of direct human mental labor for effective use. The need for and

the possibility of humanly reiterated and refined searches suggest that the human capacity for choice resists reduction to computational models or logical formulations.[1] From the perspective of automata studies, direct human choice corresponds technically to a classic sense of nondeterminism, similar to human intervention in moving a halted Turing machine to another state. Diffusion (market adoption) of techniques strongly suggests that information system consumers perceive their value (Swanson 1980).

We may have reached a practice plateau—a teleological or final stage from the theoretical considerations developed. Denying that a teleological stage has been reached would involve refuting some particularly strong theories, rejecting the account of signification derived from Saussure, denying information theory and its application to the syntactics of written language, and refuting automata theory. The strength of the individual theories is reinforced by their mutual convergence and by their closeness to real-world practice.

Conclusion

The gestalt of the computer (Rosenberg 1974), understood here in the sense of computational models and simulacra for human intelligence or capacity for choice, is further but indirectly eroded. As early as 1956, Shannon had disassociated himself from the uncritical analogical extension of information theory, while not excluding possibilities for development:

> [the] basic results of the subject [information theory] are aimed in a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences. . . . I personally believe that many of the concepts of information theory will prove useful in these other fields—and, indeed, some results are already quite promising—but the establishing of such applications is not a trivial matter of translating words to a new domain, but rather the slow tedious process of hypothesis and experimental verification. If, for example, the human being acts in some situations like an ideal decoder, this is an experimental and not a mathematical fact, and as such must be tested under a wide variety of experimental situations. (1956/1993, 462)

The decoder would act to transform a received signal into a message; this process can be both directly humanly performed (for instance, by a Morse

code telegrapher) and computationally modeled and executed. In contrast, human activity in semantic selection from a source and interpretation at the destination exceeds computational modeling. Despite Shannon's reservations—seldom directly cited—information theory became a main source for communication studies (Fiske 1990, 1), uncritically assimilating Shannon's mathematical theory of communication to Weaver's interpretation (Weaver 1949; Tidline 2004). The gestalt of the computer is acutely manifested in cognitive science, with its insistence that "theories of the mind should be expressed in a form that can be modeled in a computer program" (Johnson-Laird 1988, 52), without recourse to intuition, and in the influential linguistic modeling of the comprehension of utterances as a formal system or automaton (Sperber and Wilson 1986). Even a significant critique of cognitive science, for the absence of intentionality from computational systems, saw "no reason in principle why we couldn't give a machine the capacity to understand English, since in an important sense our bodies with our brains are precisely such machines," although not while "the operation of the machine is defined solely in terms of computational processes over formally defined elements" (Searle 1980, 422). We can detect a trace of a Cartesian inheritance in such comments: when the mind is detached from the body "the Cartesians situate the human spirit in the pineal gland, as if it were an observatory" (Vico 1710/1988, 88). Semiotic phenomena may become difficult to interpret when they are separated from physical human presence, and operation over formally differentiated elements is intrinsic to a definition of an information machine (Minsky 1967) that has not yet been supplanted (Herken 1995; Cockshott and Michaelson 2007). Drawing on the convergence of information theory and Saussurean linguistics, the argument here indicates a strong theoretical possibility that human mental labor is irreducible to computational models in the production and comprehension of written utterances.

Therefore, it is possible subtly but significantly to shift the perspective on communication and computational models and their embodiments in modern information technologies. Rather than being received as substitutes for or simulacra of human intelligence, such technologies can be regarded as the products of human mental and productive labor, building on historically accumulated understandings and technologies.

Human mental labor can be distinguished from the processes derived and abstracted from that labor that can be modeled computationally. Abstraction of processes from mental labor has intensified since the early to mid-nineteenth century, and Babbage's concern for mental labor and Boole's formalization of logic led to Turing's model of the computational process (Babbage 1963, 1989; Boole 1854; Turing 1937). With modern information technologies, transforming human mental labor into a technologically enacted process may enhance the exactness of procedures (Warner 2001, 33–46), and that exactness may emerge as both increased control over data and possibly as rigidity. A nondeterministic mode of use of systems—involving multiple direct human interventions and incorporating semantic considerations—can in some circumstances retain enhanced control without introducing rigidity.

Postscript

Introduction

In this book, "human" refers explicitly to the human sciences, the disciplinary bases in which this book is rooted. Implicit to this point, "human" also refers to modes of human being that now can be made explicit. A crucial distinction can be made for modes of being, most significantly counterpositioning being fully human against brute existence. The technological mode of information provides a secondary distinction within being fully human, a distinction made fully explicit in the preceding chapters. Thus, we now can develop a crucial question for the central transformations in being that are occurring in connection with modern information retrieval.

Modes of Being

Schematically, being fully human in a political economy and civil society can be distinguished from brute existence outside society—both as a historical transition and for individuals once civil society exists. Discourse evolves and then contributes reciprocally to being fully human. A capacity for selection is also fundamental to the contrast with brute existence (Vico 1725/2002, 33–34). Exemplified here in connection with legal codes, information retrieval developed alongside civil society and represents one part of the discourse essential to being fully human. Information retrieval can be conducted orally and also in written literate and computational modes. Selection in information retrieval represents another expression of the human capacity for selection. The contrast of full humanity with

brute existence outside society has informed this work, rather than being directly addressed within it. Contemporary changes in technological modes and information retrieval are conceived as an overlay on being fully human, not as a radical transformation that compares to the transition from brute existence; an enhancement of human capacities can be detected.

Distinctions within being fully human made by technological mode directly informed this book's discussions. We traced and distinguished orality's emergence into literacy, written literacy, and computational modes and acknowledged the transitional forms that occurred between modes; each subsequent era received, encountered, and possibly modified the inheritances from previous modes. For instance, written literacy received inheritances from orality emerging into literacy, and the computational mode received inheritances from written literacy, including its transformations of orality emerging into literacy. Thus, we understood different technological modes as transitions or additions within being fully human in civil society.

Human history is conceived here as material progress, with associated changes in consciousness. The conception is consistent with a perspective on history derived partly from Marx, which identifies

> the one element of directional change in human affairs which is observable and objective, irrespective or our subjective or contemporary wishes and value-judgments, namely the persistent and increasing capacity of the human species to control the forces of nature by means of manual and mental labour, technology and the organization of production. (Hobsbawm 1998, 41)

The reality of this directional change is:

> demonstrated by the growth of the human population of the globe throughout history, without significant set-backs, and the growth—particularly in the past few centuries—of production and productive capacity. (Hobsbawm 1998, 41)

The forces for change were understood as human physical and mental labor on the natural and humanly modified environment. Our concern here has focused directly on mental labor and its products and also on the transfer of mental labor to modern information technologies.

The labor theory of value, which can be assimilated to this conception of history, has classically been more concerned with physical or productive labor than with mental labor. By focusing on the substitution of machine

processes for human labor and some methodological preference for continuous quantitative developments, the theory can obscure the possibility of radical qualitative change. Empirically, qualitative effects could emerge from repeated quantitative substitutions, and some discussions have recognized the possibility of radical change. Marx, for instance, noted that industrial technology did not simply save or substitute for human physical labor but expanded human powers and activities:

> for, with the help of machinery, human labour performs actions and creates things which without it would be absolutely impossible of accomplishment (Marx 1858/1973, 389)

Similarly, information and communication technologies need not simply substitute for human mental labor but can extend human mental capacities. Developing considerations from the labor theory of value has the particular advantage of incorporating technology as machinery.

A parallel concern with changes in the conditions and possibilities for being in relation to mental labor can be derived from a historical perspective on communication. Specifically in relation to the history of writing and its effects on consciousness, the position formulated by Ong (1982; 1986) for the transition from orality and literacy—that writing is a technology that restructures thought—can be transformed and made both more extensive and more precise. Ong's position represents a particular instance of a more general proposition that "all new intellectual tools restructure thought" (Harris 1989), leading to a more precisely formulated and satisfying question:

> how does this innovation make possible or foster forms of thought which were previously difficult or impossible? (Harris 1989, 166)

The concern with forms of thought is analogous to the possibilities for mental labor.

A similarly precise question, connected with the conception of history as cumulative development and with the idea of additions within fully human being, can now be formulated and addressed for information retrieval: What central transformations in being are occurring in connection with modern information retrieval?

The question can be understood as a particularized instantiation and transformation of the more general question posed by Eric Hobsbawm in the conception of history as material progress:

> For, like it or not—and there are plenty of historians who don't like it—there is one central question in history which cannot be avoided, if only because we all want to know the answer to it. Namely: how did humanity get from caveman to space-traveller, from a time when we were scared by sabre-toothed tigers to a time when we are scared by nuclear explosions—that is scared not by the hazards of nature but by those we have created ourselves? (1998, 40)

The specific force for change in information retrieval has been identified throughout this book as the transfer of human labor to technology. In a transformation of Hobsbawm's question, we are concerned here with the effects of that transfer, possibly emerging from repeated quantitative substitutions.

The historical schema developed here—distinguishing orality emerging into literary, written literacy or premodernity, and computational modes or modernity—remains relevant for understanding effects on being and consciousness. The immediate focus is on the recent, still current, and continuing transition from the premodernity of written literacy to the computational mode of modernity. Effects within the professional practices of information retrieval can be distinguished from effects for users of information retrieval systems. Due to greater social diffusion and the larger number of individuals who use (rather than construct) systems, the outside effects have greater significance for collective human being.

Enhanced Selection Power

A crucial effect of recent practical developments involves the enhanced potential for recalling diverse material. While that potential has emerged from within information retrieval through repeated quantitative substitutions, we now can describe it as a qualitative transformation for users and collective being. Without the semantic labor of comprehending a complex bibliographic system, the syntactic drudgery of transcription and copying, or the physical work of searching through collections, everyday users may be approaching the condition of subject bibliographers in query formulation and selection from search results, but with some transformations and additions in their semantic search labor. Subject bibliographies were classically regarded as highly valuable or privileged forms, yielding richness without further investigative labor. The richness obtainable, both before and after current transitions, may be accompanied by uncertainty as to whether the boundaries of investigation have been reached. An enhancement has occurred in human capacities—understood as an overlay rather

than as transformation in human existence, and brought about by the continuing desire for selection power. The enhancement brings certain penalties: a change (and under certain circumstances an increase) in search labor in the use of retrieval systems, and at further remove an accentuation of the burden of the knowable.

Conclusion

The provision of enhanced selection power and its theoretical valuing provides a deep effect—the restoration of man as an artificer and the recognition of the subtlety of the processes involved in information retrieval. Rather than being subjected to a retrieval process beyond immediate control, a searcher is presented with an enhanced capacity for choice and, with certain systems, for making sets from material recalled. The new and historically unprecedented potential for enhanced forms of knowing existing textual material then can be explored productively. For instance, directly relevant to understanding language (although with effects for information retrieval), the unrivalled opportunity of full text database for exploring the semantic mutability of written word forms with different contexts now can be pursued. Rather than "Waiting for Godot [while failing] to grasp what is now within reach" (Swanson 1988, 97), we can begin to explore the potential for improving human interaction with recorded knowledge, thus becoming more fully human.

Notes

Chapter 1

1. One student's comment on the schema within the labor theoretic approach was: "It is the simplicity the schema brings to the reader that will aid them in increasing their knowledge in this area as the schema can always be remembered in its entirety due to its excellent minimalist structure." (Written comments from a student taking Information Policy, Queen's University Belfast, 2007.)

Chapter 3

1. Database management systems could be absorbed into the perspective constructed, and the commonality and contrast they have with information retrieval systems—the use of descriptions but with different schema for their construction—will be indicated. The basis for the distinction between information retrieval and database management systems also appears theoretically weak when made explicit.

The labor theoretic approach can reveal commonalities and a difference between information retrieval and database management systems (central concerns of the related subjects, although they are disjunct disciplines or communities of study, of information science and information systems (Ellis, Allen, and Wilson 1999)).

Substantial commonalities are discernible. Databases may be similarly intended to enlarge the selection power of the searcher. Selection power is also created by selection labor, disaggregated into search and description labor. Syntactic aspects of description and searching have been transferred to technology, and common technology is subject to similar possibilities and constraints of computability. A similarity in costs could be assumed.

A difference would lie in the use of different schemas for description, with database theory and practice dominated by the entity-attribute model. The model is traceable to distinctions classically made in philosophy and in formal and mathematical logic, in contrast to the emergence of description schemas for document description from more practical experience. Common, or at least strongly

Figure N3.1
Parody of entity-attribute distinction. Source: Carroll 1865/1998, 58.

analogous, concerns could be found within the different description schemas—consider, for instance, the apparent parallel between normalization of names and reduction of authors' names to canonical forms.

The distinction of attribute from entity has been questioned in philosophy and was parodied by Lewis Carroll, who had a scholarly career as the symbolic logician under his given name, Charles Dodgson (see figure N3.1).

The distinction has also been difficult to sustain in database practice.

Substantial commonalities and a significant contrast crucial to distinction between information retrieval and database management systems have been isolated and clarified.

2. I am indebted to a student for this observation.

3. W. W. Greg's comments after completing *A Bibliography of the English Printed Drama to the Restoration*, the "product of a lifetime of study," are worth recalling with regard to the prolonged labor of searching, the accumulation of expertise, the subjectively experienced limitations of that expertise in fully comprehending the topic sought, and the modification of originally held intentions though interaction with the material discovered:

lest anyone should think that looking back on my work I feel any complacency over the manner of its execution, I here admit that I can hear the caustic critic who ever sits like a familiar imp at my elbow maintaining that my problem in

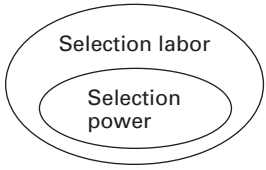
writing this introduction has been threefold: first to discover what in fact I have done, next why I did it, and lastly how best it may be defended.

The "decision to limit the work to printed plays was one of convenience or expediency" (1959, v).

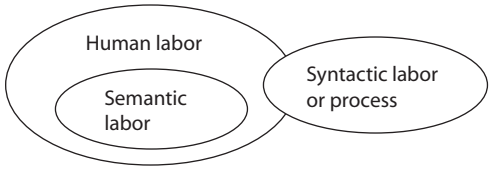
4. Some riddles informally anticipate the Boolean operators, particularly AND (consider the Anglo-Saxon examples in Hamer 1970, 95–107), but also OR and NOT.

Chapter 4

1. Selection power → Selection labor would be valid in truth-table terms, where Selection power is False and Selection labor is True; or NOT-Selection power AND Selection labor; ¬ Selection power ∧ Selection labor. Diagrammatically, the relation between Selection labor and Selection power could be represented as:



2. Semantic labor → Human labor is true for NOT-Semantic labor AND Human labor, but the number of instances likely would decrease with modern technologies, and the quantity of human labor (which is not semantic labor) would diminish accordingly. Diagrammatically, the relation between Human labor, Semantic labor, and Syntactic labor or process could be represented as:



3. The assertion sign (⊦), which can be read verbally as "it is true that …" is deliberately omitted from the assertion of the primitive terms, selection power and selection labor, and other propositions in the summary sequence. Its omission is in accord with the redundancy theory of truth itself traceable to at least the seventeenth century and Honoré Fabri (Vico 1710/1988, 145-146), which received clear modern expression by Wittgenstein:

Frege's assertion sign '⊦' is logically altogether meaningless; in Frege (and Russell) it only shows that these authors hold as true the propositions marked in this way.

'⊦' belongs therefore to the propositions no more than does the number of the proposition. A proposition cannot possibly assert of itself that it is true. (1922/1981, § 4.442)

Denying the possibility that a proposition can assert its own truth is consistent with a view of logic and technology as human constructions rather than autonomous existents and emphasizes selection power as a property of human consciousness assisted by, but not inhering in, system design.

4. The possibility of significantly differing costs or wage rates for comparable forms of direct human labor, particularly between different geopolitical regions, is acknowledged but not directly addressed.

5. Thirty four tokens of the intended type, or reference to the passage in the *Grundrisse*, were recalled, with no unintended recalled descriptions (Google 2007c).

6. Johnson did visit libraries at Oxford as part of his work on the *Dictionary of the English Language*. Some twenty years after the publication of that work, Johnson also wrote an entry for the title page of another work attributed to him: *An Account of an Attempt to Ascertain the Longitude at Sea, by an Exact Theory of the Variation of the Magnetical Needle*, "in the great Catalogue … with his own hand" (Boswell 1791/1980, 190, 194).

Chapter 5

1. A close analog to a Googlewhack (see http://www.googlewhack.com/).

2. "Direct semantic ratification" is a slightly technical term—originally from anthropology (Goody and Watt 1968, 29)—that refers, in part, to the possibility of dialog between communicators, present in interactive oral communication and restricted or delayed with written communication. This dialog is reemerging in electronic communication as still absent at the point of production of utterances but with dialogic response characteristically much more rapid than response for written communication (Warner 2004, 57–67).

3. As Ambrose Bierce noted in his *The Devil's Dictionary*.

Quotation, *n*. The act of repeating erroneously the words of another. The words erroneously repeated.

Intent on making his quotation truer,
He sought the page infallible of Brewer,
Then made a solemn vow that he would be
Condemned eternally. Ah, me, ah, me! (1906/1989, 260).

Chapter 6

1. Information theory was not available to Saussure in the early twentieth century, before its formalization in 1948 (Shannon 1948/1993), but would be "a

science which treats combination rather like algebraic equations" and would understand the linking of "sound types" (Saussure 1916/1983, 51) or other symbols in succession as transition possibilities and probabilities.

2. I am indebted to Luigina Ciolfi of the University of Limerick for informing me of this pedagogic technique.

3. For an explication and exemplification of the distinction of signifier, sign, and signified, and their role in signification when considered in relation to written language, see Warner 1994, 9–15.

Chapter 7

1. Cryptography is understood as the transformation of an identified and intercepted signal by the receiver into a message for passing to the destination. For Shannon, "[f]rom the point of view of the cryptoanalyst, a secrecy system . . . [was] almost identical with a noisy communication system" and the cryptogram was "analogous to the perturbed signal" (1949/1993, 113).

2. A search for the string "cohesive group of letters with strong" in May 2007 retrieved five documents captured by Google, seven by Google Scholar, and four by Google Book Search, with some duplication between Google and Google Scholar—a version of Shannon's own article was recalled by Google (Google 2007a, b, c).

3. Text messaging also embodies a practical, rather than theoretically informed, understanding of the possibility of reducing redundancy, while drawing on models in oral speech. Consider, for instance, the sequence *Thnx 4 infrmtn thry*.

Chapter 8

1. The analytic advantages that might follow directly for linguistics and information theory are not our primary concern here, and we will address them only summarily.

For linguistics, we have confirmed the syntagm as a primary existent and the paradigm as an analytic construct while considering computational transformations on written language, as well as the direct experience of oral and written utterances. In the priority given to language over the utterance, and in the understanding of the paradigm as a primary existent, we have detected elements of "abstract objectivism" (Vološinov 1929/1986, 79–80) in Saussure but have not rejected central Saussurean concepts. We further demonstrated the analytic value of distinguishing between syntagm and paradigm for information retrieval from full text, indicating its validity and strongly suggesting that Saussurean linguistics are relevant in information science.

The original form of information theory, including its restriction to expression, has been carefully preserved, although elements of potentially semantic selection and interpretation have emerged in Shannon's own model.

For the conjunction of linguistics and information theory, we differentiated the issues of the signified (linguistics) from the relative frequency (information theory) of sequences, if not fully substantively, then analytically. The difficulty of substantively separating issues of expression and content from frequency, and the implied increased convergence between concepts from linguistic and information theory, may reflect a common historical inheritance for practical implementation of communication that preceded theoretical abstraction, particularly for written language.

2. I am indebted to Howard Rosenbaum of Indiana University for a communication of this pedagogic technique.

3. Some commentators, notably Wittgenstein, would qualify this use of primitive:

When we have rightly introduced the logical signs, the sense of all their combinations has been already introduced with them: therefore not only "*p v q*" but also "*~(p v ~q)*", etc. etc. We should then already have introduced the effect of all possible combinations of brackets; and it would then have become clear that the proper general primitive signs are not "*p v q*", "*(∃ x) . fx*", etc., but the most general form of their combinations. (1922/1981, §5.46).

4. A student's comment on this analysis was, "The thematic concepts are powerful and systematic tools, which reach almost all the facets of the information retrieval system." (Written comments from a student taking Communicating Electronically, Queen's University Belfast, 2002.)

5. The difficulty or inability of the human mind to generate random sequences was exploited by Shannon in a machine that could anticipate human choices in calling heads or tails (1953/1993; Sloane and Wyner 1993, xvii).

6. Consider the traces of the process of production implied in Mark Twain's fictional comments in *A Connecticut Yankee in King Arthur's Court*, on receiving Morse code telegraphy in enciphered form, "and then came a click that was as familiar to me as a human voice; for Clarence had been my own pupil" (1889/1996, 479), and the synchronic access to the process of communication when intercepting signals for deciphering Enigma-produced messages in the 1939–1945 war.

7. James Murray was the original editor of the New English Dictionary, which became known as the Oxford English Dictionary (Murray 1978).

8. Figures give lightly edited selections for the opening references retrieved from Google 2005.

Chapter 9

1. A comparable development can be detected in translation systems that increasingly present selections of words for human choice (Pym 2003). The congruence becomes particularly clear if translation is conceived as choosing complexes of signifier and signified from paradigms of the target language to assign to syntagmas in the source text.

Supplementary Readings

Readings that further develop or similarly expound the approach developed in this work are limited in number by the relative novelty of the themes expounded for both the labor theoretic approach and for the application of Saussurean linguistics and Shannon's information theory to understanding retrieval from written discourse.

Some significant sources that have been adapted to the themes developed here are indicated.

The most productive way of deepening understanding of themes and concepts expounded would be dialectic exploration and testing them against a variety of widely available information systems.

Chapter 2: Selection Power and Selection Labor

For the dialogue between query transformation and selection power:

Warner, J. 2000. "In the Catalogue Ye Go for Men: Evaluation Criteria for Information Retrieval Systems." *Aslib Proceedings* 52: 76–82. Also *Information Research*. July 4, 1999. Available at: http://informationr.net/ir/4-4/paper62.html.

For a critique of the value of providing all the relevant information, placed in a fuller intellectual context:

Wilson, P. 1996b. "Interdisciplinary Research and Information Overload." *Library Trends* 45: 192–203.

Chapter 3: Description and Search Labor

For a discussion of the costs of creating descriptions and of the monetary values attached to the products of description labor:

Hayes, R. M. 2000. "Assessing the Value of a Database Company. In *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield,* ed. B. Cronin and H. B. Atkins, Medford, NJ: Information Today, Inc., 73–84.

The text of the judgment of the United Supreme Court in *Feist v. Rural* can be read.

Feist. 1991. Feist Publications, Inc. v. Rural Tel. Service Co., Inc. 499 U.S. 340.

The cited narratives involving the determination of identity can be instructively read.

Hardy, T. 1888/1976. "The Three Strangers." In *Wessex Tales*. Introduction and notes by F. B. Pinion. London: Macmillan, 13–36.

Henry, O. 1910/1993. "The Theory and the Hound." In *Selected Stories*, ed. with an introduction by G. Davenport. New York: Penguin Books, 398–406.

Chapter 5: Retrieval from Full Text

Vološinov's discussion of the single all-meaning word used by prehistoric man and its "*[m]ultiplicity of meanings*" (1986, 100–101) forms a more formal counterpart to Lewis Carroll and Humpty Dumpty:

Vološinov, V. N. 1986. *Marxism and the Philosophy of Language*. First published in 1929. Trans. by Ladislav Matejka and I. R. Titunik. New York and London, Seminar Press, 80–106 (particularly, 100–102).

Chapter 6: A Semantics for Retrieval from Full Text

Syntagmatic and associative (paradigmatic) relations as discussed by Saussure:

Saussure, F. de. 1983. *Course in General Linguistics,* Part Two, Chapters V–VII. First published 1916. C. Bally and A. Sechehaye, eds., with the collaboration of A. Riedlinger. Trans. and annotated by R. Harris. London: Duckworth, 121–136.

The pedagogic task of cutting words from a sheet of written paper (first cutting a continuous ribbon and then individual words) in order to grasp the distinction of syntagm from paradigm, is a valuable learning exercise.

For a valuable commentary on Saussure:

Harris, R. 1987. *Reading Saussure: A Critical Commentary on the Cours de Linguistique Générale*. London: Duckworth.

Chapter 7: A Syntactics for Retrieval from Full Text

In a rather neglected piece originally published in the *Encyclopedia Britannica,* Claude Shannon gives a nontechnical and widely intelligible account of information theory:

Shannon, C. E. 1993. "Information Theory." In *Claude Elwood Shannon: Collected Papers*, ed. N. J. A. Sloane and Aaron D. Wyner. First published in 1968. Piscataway, NJ: IEEE Press, 212–220.

Nontechnical readers can still profit from reading Shannon's original paper on prediction and entropy in printed English:

Shannon, C. E. 1993. Prediction and Entropy of Printed English. *Claude Elwood Shannon: Collected Papers*, ed. N. J. A. Sloane and Aaron D. Wyner. First published in 1951. Piscataway, NJ: IEEE Press, 194–208.

Chapter 8: Semantics and Syntactics for Retrieval from Full Text

Roland Barthes gives a very clear account of crucial semiotic distinctions:

Barthes, R. 1984. "Elements of Semiology." In *Writing Degree Zero & Elements of Semiology,* R. Barthes, ed. Trans. A. Lavers and C. Smith. London: Jonathan Cape, 75–172.

For an exposition of the distinction of signifier, sign, and signified:

Warner, J. 1994. *From Writing to Computers*. London: Routledge, 915.

The relation of object- to metalanguage and the significance of the costs of human description labor could be considered in relation to proposals for the Semantic Web.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. "The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities." *Scientific American*, May 2001. Available at: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

Bibliography

Amazon.com 2005. "Amazon.com Statistically Improbable Phrases." Retrieved May 2, 2005, from http://www.amazon.com/.

Amazon. 2007. "Amazon.com." Retrieved June 27, 2007, from http://www.amazon.com/.

Aristotle. *The Ethics of Aristotle: The Nicomachean Ethics*. Translated by J. A. K. Thomson, revised with notes and appendices by Hugh Tredennick, introduction and bibliography by Jonathan Barnes. London: Penguin Group, 1976. First published 323 BC.

———. *The Politics and The Constitution of Athens*. Stephen Everson, ed. Cambridge: Cambridge University Press, 1996. First published 323 BC.

Augustine. *Confessions*. Translated with an introduction by R. S. Pine-Coffin. London: Penguin Group, 1961. First published 398 BC.

Ayer, A. J. *Language, Truth and Logic*. Harmondsworth, England: Penguin Books, 1980. First published 1936 by Victor Gollancz.

Bacon, F. "Of studies." In *The Essays*. Edited with an introduction by J. Pitcher. Harmondsworth, England: Penguin Books, 1985. First published 1625, printed by Iohn Haviland for Hanna Barret and Richard Whitaker.

Babbage, C. *On the Economy of Machinery and Manufactures*, 4th edition. London: Charles Knight, 1835. In *Reprints of Economic Classics*. A. M. Kelley: New York, 1963

———. *Science and Reform: Selected Works of Charles Babbage*. Edited with an introduction by A. Hyman. Cambridge: Cambridge University Press, 1989.

Barraclough, E. D. "Online Searching in Information Retrieval," *Journal of Documentation,* Vol. 33 (1977), pp. 220–238.

Barthes, R. "Elements of Semiology." *Writing Degree Zero & Elements of Semiology,* 75–172. Translated by A. Lavers and C. Smith, London: Jonathan Cape, 1984.

Bath University Library. *Information Requirements of Researchers in the Social Sciences,* R. Barthes, ed. Bath: Bath University Library, 1971.

Belkin, N. J. and A. Vickery. *Interaction in Information Systems: A Review of Research from Document Retrieval to Knowledge-Based Systems (Library and Information Research Report 35)*. London: British Library, 1985.

Bell, E. T. *Men of Mathematics*. London: Penguin Books, 1937.

Berger, T. "Lossy Source Coding." In *Information Theory: 50 Years of Discovery,* edited by S. Verdú and S. W. McLaughlin, 649–679. Piscataway: Wiley-IEEE Press, 2000.

Bergin, T. G. and M. H. Fisch. "Introduction." In G. Vico. *The New Science of Giambattista Vico*. Unabridged translation of the 3rd edition (1744) with the addition of "Practice of the New Science," xix–xlv. Translated by T. G. Bergin and M. H. Fisch. Ithaca, NY: Cornell University Press, 1976.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. "The Semantic Web: A New Form of Web Content that is Meaningful to Computers Will Unleash a Revolution of New Possibilities." *Scientific American*, (May 2001). Available at: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

Bierce, A. *The Enlarged Devil's Dictionary*. Researched and edited by E. J. Hopkins. Preface by J. M. Myers. London: Penguin Books, 1989. First published 1906 by Doubleday, Page, & Company as *The Cynics's Word Book*.

Blair, D. C. "Knowledge Management: Hype, Hope, or Help?" *Journal of the American Society for Information Science and Technology,* Vol. 53 (2002), pp. 1019–1028.

Bochenski, I. M. *History of Formal Logic*. Translated by I. Thomas. Indiana: Notre Dame, 1961.

Boole, G. *An Investigation of the Laws of Thought. On Which are Founded the Mathematical Theories of Logic and Probabilities*. London: Walton and Maberly, 1854.

Boolos, G. S. and Jeffrey, R. C. *Computability and Logic*, 3rd edition. Cambridge: Cambridge University Press, 1989.

Borges, J. L. *The Book of Imaginary Beings*. J. L. Borges with M. Guerrero. Revised, enlarged and translated by N. T. di Giovanni in collaboration with the author. London: Penguin Books, 1974.

Boswell, J. *Life of Johnson*. Edited by R. W. Chapman and revised by J. D. Fleeman. Introduction by P. Rogers. Oxford: Oxford University Press, 1980. First published 1791, printed by Henry Baldwin, for Charles Dilly.

Bradford, S. C. *Documentation,* 2nd edition. High Wycombe, Bucks: University Microfilms for the College of Librarianship Wales, 1971. First published 1948 by C. Lockwood.

Buckland, M. "Information as Thing," *Journal of the American Society for Information Science,* Vol. 42 (1991), pp. 351–360.

Buckland, M. and C. Plaunt. "On the Construction of Selection Systems," *Library Hi Tech,* Vol. 12 (1994), pp. 15–28.

Carroll, L. "Alice's Adventures in Wonderland." In L. Carroll, *Alice's Adventures in Wonderland and Through the Looking-Glass and What Alice Found There,* edited with an introduction and notes by H. Haughton, 1–110. London: Penguin Books, 1998. First published 1865 by Macmillan and Co..

———. "Through the Looking Glass, and What Alice Found There." In L. Carroll, *Adventures in Wonderland and Through the Looking-Glass and What Alice Found There,* edited with an introduction and notes by H. Haughton, 111–246. London: Penguin Books, 1998. First published 1872 by Macmillan and Co.

Chandler, R. *Raymond Chandler Speaking*. D. Gardiner and K. S. Walker, eds. Introduction by P. Skenazy. Berkeley: University of California Press, 1997. First published 1962 by Houghton Mifflin.

Cherry, C. *On Human Communication: A Review, A survey, and A Criticism*. 3rd ed. Cambridge, Mass: MIT Press, 1978.

Childe, V. G. *Society and Knowledge*. London: George Allen & Unwin, 1956.

Cleverdon, C. W. *ASLIB Cranfield Research Project: Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Cranfield: College of Aeronautics, 1962.

Cleverdon, C., J. Mills, and M. Keen. *Factors Determining the Performance of Indexing Systems*. Cranfield: College of Aeronautics, 1966.

Cockshott, P. and G. Michaelson. "Are There New Models of Computation? Reply to Wegner and Eberbach," *The Computer Journal,* Vol. 50 (2007), pp. 232–247.

Collins, R. *The Sociology of Philosophies: A Global Theory of Intellectual Change*. Cambridge, Mass: Belknap Press of Harvard University Press, 1998.

Copyright Decisions. *Decisions of the United States Courts Involving Copyright and Literary Property 1789–1909*. Washington: Copyright Office, Library of Congress, 1909.

Culler, J. *Saussure*. London: Fontana Press, 1988.

Darwin, C. *The Origin of Species by Means of Natural Selection or The Preservation of Favoured Races in the Struggle for Life*. London: Penguin Group, 1968. First published 1859 by John Murray.

Davis, M. "Influences of Mathematical Logic on Computer Science." In *The Universal Turing Machine: A Half Century Survey,* edited by R. Herken, 315–326. Oxford: Oxford University Press, 1988.

Davis, M. *The Universal Computer: The Road from Leibniz to Turing*. New York: W. W. Norton, 2000.

Dickens, C. *The Personal History of David Copperfield*. Oxford: Oxford University Press, 1981. First published 1850 by Bradbury & Evans.

———. *Hard Times*. Oxford and New York: Oxford University Press, 1989. First published 1854 by Bradbury & Evans.

———. *Great Expectations*. London: Cumberlege, 1946. First published 1861 by Chapman and Hall.

Dijkstra, E. W. "Letters to the Editor: Go To Statement Considered Harmful (Letters to the Editor)," *Communications of the ACM,* Vol. 11 (1968), pp.147–148.

Donovan, E. *The Natural History of British Quadrapeds; Consisting of Coloured Figures, Accompanied with Scientific and General Descriptions, of All the Species That Are Known to Inhabit the British Isles . . . and Also Such as Are Clearly Authenticated to Have Been Originally Indigenous, But Are Now Extirpated, or Become Extremely Rare; the Whole Arranged in Systematic Order, After the Manner of Linnaeus*. London: Printed for the Author, and for F. C. and J. Rivington, 1820.

Ellis, D. "Theory and Explanation in Information Retrieval Research," *Journal of Information Science,* Vol. 8 (1984), pp. 25–38.

———. "The Physical and Cognitive Paradigms in Information Retrieval Research," *Journal of Documentation,* Vol. 48 (1992), pp. 45–64.

———. *Progress and Problems in Information Retrieval*. London: Library Association, 1996.

Ellis, D., D. Allen, T. Wilson. "Information Science and Information Systems: Conjunct Subjects Disjunct Disciplines," *Journal of the American Society for Information Science,* Vol. 50 (1999), pp. 1095–1107.

Ellis, D. and A. Vasconcelos. "Ranganathan and the Net: Using Facet Analysis to Search and Organise the World Wide Web," *Aslib Proceedings,* Vol. 51 (1999), pp. 3–10.

Feist Publications, Inc. v. Rural Telephone Service Co., Inc., 499 U.S. 340 (1991).

Fiske, J. *Introduction to Communication Studies*. 2nd edition. London: Routledge, 1990.

Freud, S. *Jokes and Their Relation to the Unconscious*. Translated and edited by J. Strachey and A. Richards. Harmondsworth, England: Penguin Books, 1976. First published 1905 by Deuticke.

Gardin, J-C. "Document Analysis and Linguistic Theory," *Journal of Documentation,* Vol . 29 (1973), pp. 137–168.

Gardner, M. *Logic Machines and Diagrams*, 2nd edition. Brighton: Harvester, 1983.

Goody, J. and I. Watt, "The Consequences of Literacy." In *Literacy in Traditional Societies, edited by J. Goody,* 27–68. Cambridge: Cambridge University Press, 1968.

Google. 2004. Google Advanced Search. Retrieved December 17, 2004, from http://www.google.co.uk/advanced_search?hl=en.

———. 2005. Google Advanced Search. Retrieved May 5, 2005, from http://www.google.co.uk/advanced_search?hl=en.

———. 2007a. Google. Retrieved June 27, 2007, from http://www.google.co.uk/advanced_search?hl=en.

———. 2007b. Google Advanced Image Search. Retrieved June 27, 2007, from http://images.google.com/advanced_image_search?hl=en.

———2007c. Google Book Search. Retrieved July 25, 2007, from http://books.google.co.uk/advanced_book_search

———. 2008. Google Advanced Scholar. Retrieved June 2, 2008, from http://scholar.google.com/advanced_scholar_search

Gosden, C. *Prehistory: A Very Short Introduction*. Oxford: Oxford University Press, 2003.

Greg, W. W. *A Bibliography of the English Printed Drama to the Restoration*, Volume IV. London: printed for the Bibliographical Society at the University Press, Oxford, 1959.

Gregory, V. L. "Review of J. Warner. Humanizing Information Technology," *Library Resources & Technical Services,* Vol. 49 (2005), pp. 59–60.

Griesbaum, J. 2000. Evaluierung Hybrider Suchsysteme im WWW. iDiplomarbeit, Universität Konstanz Informationswissenschaft [Evaluation of Hybrid Internet Search Systems]. Retrieved February 21, 2007, from http://www.inf.uni-konstanz.de/~griesbau/files/evaluierung_hybrider_suchsysteme_im_www.pdf.

Hardy, T. "The Three Strangers." In *Wessex Tales*, 13–36. Introduction and notes by F. B. Pinion. London: Macmillan, 1976. First published 1888 by MacMillan.

Hamer, R. *A Choice of Anglo-Saxon Verse*. London: Faber and Faber, 1970.

Harris, R. *Reading Saussure: A Critical Commentary on the Cours de Linguistique Générale*. London: Duckworth, 1987.

———. "How does writing restructure thought?" *Language and Communication,* Vol. 9 (1989), pp. 99–106.

———. *Signs of Writing*. Routledge: London and New York, 1995.

Hayes, R. M. "Assessing the Value of a Database Company." *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield,* edited by B. Cronin and H. B. Atkins, 73–84. Medford, NJ: Information Today, Inc., 2000.

Heine, M. H. "The 'Question' as a Fundamental Variable in Information Science." *Theory and Application of Information Research (Proceedings of the Second International Research Forum on Information Science*, 3–6 August 1977, Royal School of Librarianship, Copenhagen), edited by O. Harbo and L. Kajberg, 137–145. London: Mansell, 1980.

Henry, O. "The Theory and the Hound." In *Selected Stories*, edited with an introduction by G. Davenport, 398–406. New York: Penguin Books, 1993. First published 1910 by Doubleday, Page & Co..

Herken, R., ed. *The Universal Turing Machine: A Half-Century Survey*, 2nd edition. Wien and New York: Springer-Verlag, 1995.

Johnson, S. "Preface to the Dictionary." In *Johnson's Dictionary: A Modern Selection*, edited by E. L. McAdam and G. Milne, 3–29. London and Basingstoke: Macmillan, 1982a. First published 1755, printed by W. Strahan, for J. and P. Knapton [etc.].

———. *Johnson's Dictionary: A Modern Selection*, edited by E. L. McAdam and G. Milne, eds. London and Basingstoke: Macmillan, 1982b. First published 1755, printed by W. Strahan, for J. and P. Knapton [etc.].

———. *A Dictionary of the English Language [Electronic Resource]: on CD-ROM*. Ed. A. McDermott. Cambridge: Cambridge University Press, 1996.

Johnson-Laird, P. N. *The Computer and the Mind: An Introduction to Cognitive Science*. London: Fontana, 1988.

Linné, C. *The Animal Kingdom, or Zoological System, of the Celebrated Sir Charles Linnaeus. Class I. Mammalia: Containing a Complete Systematic Description, Arrangement, and Nomenclature, of All the Known Species and Varieties. . . .* Edinburgh: Printed for A. Strahan, T. Cadell, 1792.

Liversidge, A. "Profile of Claude Shannon." In *Claude Elwood Shannon: Collected Papers,* edited by N. J. A. Sloane and Aaron D. Wyner, xix–xxxiii. Piscataway, NJ: IEEE Press, 1993.

Lyotard, J-F. *The Postmodern Condition: a Report on Knowledge*. Translated by G. Bennington and B. Massumi. Minneapolis: University of Minnesota Press, 1984.

MacKenzie, D. *Mechanizing Proof: Computing, Risk, and Trust*. Cambridge, Mass: MIT Press, 2001.

Marx, K. "The Eighteenth Brumaire of Louis Bonaparte." In K. Marx. *Surveys from Exile*, edited and introduced by David Fernbach, 143–249. Harmondsworth, Eng: Penguin Books in association with New Left Review, 1973. First published 1852 in *Die Revolution*.

———. *Grundrisse: Foundations of the Critique of Political Economy (Rough Draft)*. Translated with a foreword by Martin Nicolaus. London: Penguin Books in association with New Left Review, 1973. Written c.1858 and first published 1939 by Verlag für Fremdsprachige Literatur.

———. *Capital: A Critique of Political Economy*, Volume One. Introduced by E. Mandel and translated by B. Fowkes. London and New York: Penguin Books in association with New Left Review, 1976. First published 1867 by O. Meissner and L. W. Schmidt.

———. *Capital: A Critique of Political Economy*, Volume Three. Introduced by E. Mandel and translated by B. Fowkes. London and New York: Penguin Books in association with New Left Review, 1981. First published 1894 by Meissner.

McKenzie, D. F. "Speech-Manuscript-Print," *The Library Chronicle of the University of Texas,* Vol. 20 (1990), pp. 87–109.

McLeod, N. S., J. Perelman, and W. B Johnston. *Horse Feathers*. Hollywood: Paramount, 1932.

Minsky, M. *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: Prentice-Hall, 1967.

Montgomery, C. A. "Linguistics and Information Science," *Journal of the American Society for Information Science,* Vol. 23 (1972), pp. 195–219.

Moradi, H., J. W. Grsymala-Busse, and J. A. Roberts. "Entropy of English Text: Experiments with Humans and a Machine Learning System Based on Rough Sets," *Information Sciences,* Vol. 104 (1998), pp. 31–47.

Morrison, D. "Index to *The Times*," *Journal of Documentation,* Vol. 42 (1986), pp. 189–195.

Murra, K. O. "History of Some Attempts to Organize Bibliography Internationally." In *Bibliographic Organization: Papers Presented Before the Fifteenth Annual Conference of the Graduate Library School July 24–29, 1950,* edited by J. H. Shera and M. E. Egan 24–53. Chicago: University of Chicago Press, 1951.

Murray, K. M. E. *Caught in the Web of Words: James A. H. Murray and the Oxford English Dictionary*. New Haven: Yale University Press, 1978.

Njal. *Njal's Saga*. Translated with an introduction by M. Magnusson and H. Pálsson. Harmondsworth: Penguin Books, 1960. First published 1280.

Ong, W. J. *Orality and Literacy: the Technologizing of the Word*. London and New York: Methuen, 1982.

———. "Writing is a Technology That Restructures Thought." In *The Written Word: Literacy in Transition (Wolfson College Lectures 1985*), edited by G. Baumann, 23–50. Clarendon Press: Oxford, 1986.

Orwell, G. "Why I Write." In G. Orwell. *Collected Essays, Journalism and Letters,* Volume I, edited by S. Orwell and I. Angus, 23–30. Harmondsworth, England: Penguin Books, 1970. First published 1946 in *Gangrel*.

Oxford Dictionaries. *Oxford English Dictionary*, C. Soanes and A. Stevenson, eds. Oxford University Press, 2005. Also available at http://dictionary.oed.com.

Palmer, S. *Palmer's Index to The Times: Spring Quarter*. Richmond House, Shepperton on Thames: Samuel Palmer, 1885.

Pierce, J. R. *An Introduction to Information Theory: Symbols, Signals and Noise*, 2nd edition. New York: Dover Publications, 1980.

Plutarch. *The Rise and Fall of Athens: Nine Greek Lives by Plutarch*. Translated with an introduction by I. Scott-Kilvert. Harmondsworth, England: Penguin Books, 1960. First published in 100 BC.

Pow, T. "Did Englishmen Eat Each Other?: In the 150th Anniversary Year of Sir John Franklin's Death, Tom Pow Considers Why His Failed Endeavour Overshadowed the Greater Achievements of the Orcadian Explorer John Rae," *Scotland,* January 26, 1997, p. 8.

Powers, T. 2005. "Black Arts," *New York Review of Books*, May 12, 2005, pp. 21–25.

Pym, A. "Redefining Translation Competence in an Electronic Age," *Meta,* Vol. 48 (2003), pp. 481-497.

Quine, W. V. O. "New Foundations for Mathematical Logic." In W. V. O. Quine. *From a Logical Point of View: 9 Logico-Philosophical Essays.* Cambridge: Harvard University Press, 1953. First published in 1937.

Railton, S. 2005. "Mark Twain in His Times". Retrieved March 17, 2005, from http://etext.lib.virginia.edu/railton/index2.html.

Ramsey, F. P. "The Foundations of Mathematics In F. P. Ramsey. *Philosophical Papers,* edited by D. H. Mellor, 164–224. Cambridge: Cambridge University Press, 1990. First published 1925 in *Proceedings of the London Mathematical Society*.

Roberts, N. "Social Considerations Towards a Definition of Information Science," *Journal of Documentation* Vol. 32 (1976), pp. 249–257.

———. "Communication and the Bibliographical System of the Social Sciences." In *Use of Social Sciences Literature,* edited by N. Roberts, 1–27. London: Butterworths, 1977.

———. "Online: In an Educational Cul-de-Sac?" *Education for Information,* Vol. 7 (1989), pp. 101–106.

Rosenberg, V. 1974. "The Scientific Premises of Information Science," *Journal of the American Society for Information Science,* Vol. 25 (1974), pp. 263–269.

Saussure, F. de. *Course in General Linguistics*. Edited by C. Bally and A. Sechehaye with the collaboration of A. Riedlinger. Translated and annotated by R. Harris. London: Duckworth, 1983. First published 1916 by Payot.

Searle, J. R. "Minds, Brains and Programs," *The Behavioral and Brain Sciences,* Vol. 3 (1980), pp. 417–457.

Shakespeare, W. *Macbeth. The Arden Edition of the Works of William Shakespeare* edited by K. Muir. London and New York: Routledge, 1988.

Shannon, C. E. "A Mathematical Theory of Communication." In *Claude Elwood Shannon: Collected Papers,* edited by N. J. A. Sloane and A. D. Wyner, 5–83. Piscataway: IEEE Press, 1993. First published 1948 *Bell System Technical Journal*.

———. "Communication Theory of Secrecy Systems." In *Claude Elwood Shannon: Collected Papers,* edited by N. J. A. Sloane and A. D. Wyner, 84–143. Piscataway: IEEE Press, 1993. First published 1949 in *Bell System Technical Journal*.

———. "Prediction and entropy of printed English." In *Claude Elwood Shannon: Collected Papers,* edited by N. J. A. Sloane and A. D. Wyner, 194–208. Piscataway, NJ: IEEE Press, 1993. First published 1951 in *Bell System Technical Journal*.

———. "A Mind-Reading (?) Machine." In *Claude Elwood Shannon: Collected Papers,* edited by N. J. A. Sloane and A. D. Wyner, 688–689. Piscataway, NJ: IEEE Press, 1993.

———. "The Bandwagon." In *Claude Elwood Shannon: Collected Papers,* edited by N. J. A. Sloane and A. D. Wyner, 462. Piscataway: IEEE Press, 1993. First published 1956 in *Institute of Radio Engineers, Transactions on Information Theory*.

———. "Information theory." In *Claude Elwood Shannon: Collected papers,* edited by N. J. A. Sloane and A. D. Wyner, 212–220. Piscataway, NJ: IEEE Press, 1993. First published 1968 in *Encyclopedia Britannica*.

Shera, J. H. "Foundations of a Theory of Bibliography." In J. H. Shera. *Libraries and the Organization of Knowledge*, edited by D. J. Foskett, 18–33. Hamden, Connecticut: Archon Books, 1965.

———. *Social Epistemology, General Semantics, and Librarianship*. D. J. Foskett, ed., 12–17. Hamden, Connecticut: Archon Books, 1965.

Shneiderman, B. *Leonardo's Laptop: Human Needs and the New Computing Technologies*. Cambridge, Mass: MIT Press, 2003.

Short, W. R. 2008. "Norse Laws and Legal Procedures." Retrieved January 15, 2008 from, http://www.hurstwic.org/history/articles/society/text/laws.htm.

Sloane, N. J. A. and A. D. Wyner. "Biography of Claude Elwood Shannon." In *Claude Elwood Shannon: Collected Papers,* edited by N. J. A. Sloane and A. D. Wyner, xi–xvii. Piscataway: IEEE Press, 1993.

Smithson, S. "Information Retrieval Evaluation in Practice: A Case Study Approach," *Information Processing and Management,* Vol. 30 (1994), pp. 205–221.

Sparck Jones, K. and M. Kay. *Linguistics and Information Science*. New York and London: Academic Press, 1973.

Sperber, D. and D. Wilson. *Relevance: Communication and Cognition*. Oxford: Basil Blackwell, 1986.

Starobinski, J. *Words Upon Words: The Anagrams of Ferdinand de Saussure*. Translated by O. Emmet. New Haven and London: Yale University Press, 1979.

Stevens, A. *Ariadne's Clue: A Guide to the Symbols of Mankind*. London: Allen Lane The Penguin Press, 1998.

Swanson, D. R. 1980. "Libraries and the Growth of Knowledge." In *The Role of Libraries in the Growth of Knowledge*, edited by D. R. Swanson, 112–136. Chicago: University of Chicago Press.

———. 1988. "Historical Note: Information Retrieval and the Future of an Illusion," *Journal of the American Society for Information Science,* Vol. 39 (1988), pp. 92–98.

Tidline, T. J. "Working in Shannon's Shadow: Mistaken Identity and Persistent Entropy of Information Concepts." Poster presented at the Annual Meeting of the American Society for Information Science and Technology, Providence, Rhode Island, November 2004.

Times 1885. *The Times (London)*. 1885.

Turing, A. M. "On Computable Numbers, With an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society,* Vol. 42 (1937), pp. 230–265.

Twain, M. *A Connecticut Yankee in King Arthur's Court*. First published 1889. New York and Oxford: Oxford University Press, 1996.

UNESCO/Library of Congress Bibliographic Survey. *Bibliographical Services, Their Present State and Possibilities of Improvement. Report Prepared as a Working Paper for an International Conference on Bibliography*. Washington, 1950.

Van Rijsbergen, C. J. *Information Retrieval*, 2nd edition. London: Butterworth-Heinemann, 1979.

Verdú, S. and S. W. McLaughlin. *Information Theory: 50 Years of Discovery*. New York: IEEE Press, 2000.

Verne, J. *Backwards to Britain*. Translated by J. Valls-Russell. Edinburgh and New York: Chambers, 1992. First published as *Voyage à Reculons en Angleterre et en Ecosse*, 1989.

Vico, G. *On the Most Ancient Wisdom of the Italians: Unearthed from the Origins of the Latin Language*.. Translated with introduction and notes by L. M. Palmer. Ithaca: Cornell University Press, 1988. First published 1710, Ex Typographia Felicis Mosca.

———. *The First New Science*. Edited and translated by L. Pompa. Cambridge: Cambridge University Press, 2002. First published 1725, Per F. Mosca.

———. The *New Science of Giambattista Vico*. Unabridged translation of the third edition (1744) with the addition of "Practice of the New Science" Translated by T. G. Bergin and M. H. Fisch. Ithaca, NY and London: Cornell University Press, 1976. First published 1744, Nella stamperia Muziana, a spese di G. e S. Elia.

———. *The Autobiography of Giambattista Vico*. Translated from the Italian by M. H. Fisch and T. G. Bergin. Ithaca: Cornell University Press, 1990. First published 1818, Presso Porcelli.

Vološinov, V. N. *Marxism and the Philosophy of Language*. Translated by L. Matejka and I. R. Titunik. New York and London: Seminar Press, 1986. First published 1929 by Priboi.

Warner, J. "Writing and Literary Work in Copyright: A Binational and Historical Analysis," *Journal of the American Society for Information Science,* Vol. 44 (1993), pp. 307–321.

———. *From Writing to Computers*. London and New York: Routledge, 1994.

———. "The Public Reception of the Research Assessment Exercise 1996," *Aslib Proceedings,* Vol. 49 (1997), pp. 263–276.

———. 2000. "In the Catalogue Ye Go for Men: Evaluation Criteria for Information Retrieval Systems," *Aslib Proceedings,* Vol. 52 (2000), pp. 76–82. Also from *Information Research*, Vol. 4 (July 1999). Available at http://informationr.net/ir/4-4/paper62.html.

———. *Information, Knowledge, Text*. Lanham, MD: Scarecrow Press, 2001.

———. "Information and Redundancy in the Legend of Theseus," *Journal of Documentation,* Vol. 59 (2003), pp. 540–557.

———. *Humanizing Information Technology*. Lanham, MD: Scarecrow Press, 2004.

———. "Labor in Information Systems," *Annual Review of Information Science and Technology,* Vol. 39 (2005a), pp. 551–573.

———. "An Information Dynamic: Technologies for the Reproduction of Written Utterances," *Aslib Proceedings,* Vol. 57 (2005b), pp. 412–423.

Weaver, W. "Recent Contributions to the Mathematical Theory of Communication." In *The Mathematical Theory of Communication,* edited by C. E. Shannon and W. Weaver, 1–28. Urbana: University of Illinois Press, 1949.

Webster, F. *Theories of the Information Society*, 2nd edition. London and New York: Routledge, 2002.

Weisman, R. "Talking Search Technology: Eric Schmidt, Chairman and CEO, Google Inc.," *Boston Globe*, February 4, 2002, p. D1.

Wiener, N. *The Human Use of Human Beings: Cybernetics and Society*. Revised edition. New York: De Capo Press, 1954.

Wilkins, J. *An Essay Towards a Real Character and a Philosophical Language*. Menston, England: Scolar Press, 1968. First published in 1668.

Wilson, P. 1968. *Two Kinds of Power: An Essay on Bibliographical Control*. Berkeley and Los Angeles: University of California Press, 1968.

———. "Situational Relevance," *Information Storage and Retrieval,* Vol. 9 (1973), pp. 457–471.

———. "Review of F. Webster. Theories of the Information Society," *College and Research Libraries,* Vol. 57 (1996a), pp. 487–489.

———. "Interdisciplinary Research and Information Overload," *Library Trends,* Vol. 45 (1996b.), pp. 192–203.

———. "Some Consequences of Information Overload and Rapid Conceptual Change." In *Information Science: From the Development of the Discipline to Social Interaction*, edited by J. Olaisen, E. Munch-Petersen, and P. Wilson, 21–34. Oslo: Scandinavian University Press, 1996c.

———. "Review of E. Svenonius. The Intellectual Foundations of Information Organization," *College and Research Libraries,* Vol. 62 (2001), pp. 203–204.

Wittgenstein, L. *Tractatus Logico-Philosophicus*. London and New York: Routledge and Kegan Paul, 1981. First published 1922 by Routledge & Kegan Paul.

WorldCat. 2007. WorldCat. OCLC. Retrieved June 27, 2007, from http://firstsearch.uk.oclc.org/.

Zipf, G. K. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. London: George Routledge, 1936.

Index

Note: *b* stands for box; *f* for figure; *n* for note; *t* for table.